Statistics Fundamentals

Succinctly®

by Katharine Alexis Kormanik

Syncfusion | Technology Resource Portal

Statistics Fundamentals

Succinctly

By
Katharine Alexis Kormanik

Foreword by Daniel Jebaraj

2501 Aerial Center Parkway
Suite 200
Morrisville, NC 27560
USA

Technical Reviewer: Joe Booth
Copy Editor: John Elderkin
Acquisitions Coordinator: Hillary Bowling, online marketing manager, Syncfusion, Inc.
Proofreader: Tres Watkins, content development manager, Syncfusion, Inc.

Table of Contents

The Story behind the *Succinctly* Series of Books 6

About the Author 8

Introduction 9

About *Foundational and Computational Statistics Succinctly* 9
Get set up 10

Chapter 1 Central Tendency 12

Calculate measures of center 12

Chapter 2 Variability 18

Calculate measures of spread 18

Chapter 3 Distributions 25

Visualize the shape of data 25

Chapter 4 Standardizing 31

Use distributions to find probabilities 31
Determine what is significantly unlikely 35

Chapter 5 One-Sample Z-Test 38

Calculate the likelihood of a random sample 38
Find a range for the true mean 43
Calculate the likelihood of a proportion 46

Chapter 6 T-Tests 48

Hypothesis test when population parameters are unknown 48

Chapter 7 ANOVA 59

Test for differences between three or more samples 59
Compare samples based on one factor 59
Compare samples based on two factors 70
Additional ANOVA information 76

Chapter 8 Tabulated Data 77

Test for significance with tabulated data 77

Chapter 9 Linear Regression 81

Predict one variable with another 81

Afterward 90

Continue your statistics journey 90

Glossary 91

Appendix 96

z-table 98
t-table 1 99
t-table 2 100
f-table 1 101
f-table 2 102
χ^2 table 1 103
χ^2 table 2 104

The Story behind the *Succinctly* Series of Books

Daniel Jebaraj, Vice President
Syncfusion, Inc.

Staying on the cutting edge

As many of you may know, Syncfusion is a provider of software components for the Microsoft platform. This puts us in the exciting but challenging position of always being on the cutting edge.

Whenever platforms or tools are shipping out of Microsoft, which seems to be about every other week these days, we have to educate ourselves, quickly.

Information is plentiful but harder to digest

In reality, this translates into a lot of book orders, blog searches, and Twitter scans.

While more information is becoming available on the Internet and more and more books are being published, even on topics that are relatively new, one aspect that continues to inhibit us is the inability to find concise technology overview books.

We are usually faced with two options: read several 500+ page books or scour the web for relevant blog posts and other articles. Just as everyone else who has a job to do and customers to serve, we find this quite frustrating.

The *Succinctly* series

This frustration translated into a deep desire to produce a series of concise technical books that would be targeted at developers working on the Microsoft platform.

We firmly believe, given the background knowledge such developers have, that most topics can be translated into books that are between 50 and 100 pages.

This is exactly what we resolved to accomplish with the *Succinctly* series. Isn't everything wonderful born out of a deep desire to change things for the better?

The best authors, the best content

Each author was carefully chosen from a pool of talented experts who shared our vision. The book you now hold in your hands, and the others available in this series, are a result of the authors' tireless work. You will find original content that is guaranteed to get you up and running in about the time it takes to drink a few cups of coffee.

Free forever

Syncfusion will be working to produce books on several topics. The books will always be free. Any updates we publish will also be free.

Free? What is the catch?

There is no catch here. Syncfusion has a vested interest in this effort.

As a component vendor, our unique claim has always been that we offer deeper and broader frameworks than anyone else on the market. Developer education greatly helps us market and sell against competing vendors who promise to “enable AJAX support with one click,” or “turn the moon to cheese!”

Let us know what you think

If you have any topics of interest, thoughts, or feedback, please feel free to send them to us at succinctly-series@syncfusion.com.

We sincerely hope you enjoy reading this book and that it helps you better understand the topic of study. Thank you for reading.

Please follow us on Twitter and “Like” us on Facebook to help us spread the word about the *Succinctly* series!

About the Author

Katharine Alexis Kormanik began tutoring math at age 13. By the time she graduated college (where she majored in math and economics), she had taught more than 100 students. After completing a master's degree in International Comparative Education at Stanford University, she consulted for a number of ed-tech companies, contributed to several digital mathematics textbooks, and designed and taught Udacity's online statistics courses *Intro to Descriptive Statistics* and *Intro to Inferential Statistics.* Following her stint at Udacity, she continued her work as an online educator by designing MOOCs for Stanford Graduate School of Business, McKinsey Academy, African Leadership University, and Dribbble.

She is passionate about creating engaging learning experiences. Her work involves coordinating with subject-matter experts (faculty at leading institutions, partners at major corporations), videographers, UX-UI designers, graphic designers, and engineers to design effective courses and bring them to life. When she has time, she continues to create educational math materials such as this *Succinctly* book, and she offers short algebra courses—specifically, algebra concepts that are critical to understand calculus—at turnthewheel.thinkific.com.

Introduction

About *Foundational and Computational Statistics Succinctly*

In today's world, analyzing large data sets is more important than ever. Society's shift into the digital sphere has resulted in hundreds of thousands of data points being right at our fingertips. With this shift, statistical analysis holds greater relevance than ever, and people who understand how to work with numbers have a highly valued and sought-after skill. Why? Because numbers tell a story, and this story enables us to make the best decisions.

This e-book on foundational (i.e. the theories behind the analyses) and computational (i.e. actually performing the analyses) statistics covers visualizing and describing data, making conjectures about **populations** (the entirety of the subjects that make up a particular group) based on **samples** (a group of subjects from the population), and using statistical tests to determine if two or more samples or populations are significantly different.

In each lesson, you'll see key statistics terms noted in bold upon first usage. These terms are defined in the glossary at the end of the e-book.

In each chapter, you'll build a solid foundation of the theory and methodology behind each statistical procedure, and you'll explore real-world examples in which you might use them. Many of the examples here are simplified for the sake of demonstrating each statistical procedure, but I will also point out additional considerations in order to ensure your tests are robust.

You'll also learn how to perform basic analysis in the statistical program R with a large sample data set and learn how to interpret the results. R is a free, open source project that contains a language and environment for statistical computing and graphics, and it is one of the most widely used programs for doing analysis.

This e-book is a condensed and accelerated version of *Street-Smart Stats: A Friendly Introduction to Statistical Research Methods*. It assumes an intermediate knowledge of algebra and covers more advanced statistical analyses (e.g., ANOVA, multiple regression) in depth.

I would love to hear your feedback on the book. Feel free to contact me with questions or comments anytime at www.turnthewheel.org/about.

Get set up

You'll see the R codes in each chapter, but to practice them yourself you'll need to download R and input the data sets we analyze. To download R, visit http://cran.r-project.org/ and click the link that applies to your operating system. (To learn more about R, visit https://www.r-project.org/.)

Practice using R with real data

We'll be working with one large data set downloaded from the National Center for Education Statistics (NCES) website throughout this e-book. The data includes select variables from the Education Longitudinal Study (ELS) of 2002-2012 (demographic variables, standardized test scores, student activities, hours spent on homework, and 2011 family income).

The original NCES data for the study consisted of more than 16,000 students. Assuming these students were selected randomly, this sample can be used to draw conclusions about the entire population of American students.

However, the data set you will download and input into R excludes students with any missing values for the selected variables, which cuts the **sample size (n)** down to 8,247.

You can practice downloading and inputting the data set into R by visiting my site http://turnthewheel.org /street-smart-stats/afterward/ and clicking the first link under Resources: "ELS Longitudinal Study 2002-2012." The data opens in a Google spreadsheet.

File > Download as > Comma-separated values (.csv, current sheet)

After downloading the file as a .csv, rename it "ELS2002.csv" and save it to your working directory so R can access it. If you don't know your working directory, type **`getwd()`** in the R console. R will output the folder name in which you should save all your .csv files.

Once the file is in your working directory, you will input the data into R using the **`read.csv()`** function.

The **`head()`** function is optional; it gives you each variable name as well as the first six values. This is a quick way to see all the variables in your data set and the types of values in each (e.g., many of the variables are binary, with values 0 or 1 representing "No" and "Yes," while other variables are continuous). You can read a description of each variable in this data set in the Appendix.

The **`attach()`** function allows R to recognize variable names so that you can analyze each (e.g., you can find the average test score).

The following Code Listing shows the R inputs and outputs we have covered thus far.

Code Listing 1

```
> els2002 = read.csv(file = "ELS2002.csv", head = TRUE, sep = ",")
  # allows R to "read" the data set, and specifies that there is a header
  (head = TRUE) and that the values are separated by commas (sep = ",")

> head(els2002) #outputs the name and first six values of each variable

  gender race   ses  test homework tv_games work grades
1      0    0 -0.40 47.37        0        2    0      0
2      0    0 -0.73 39.65        0        4    0      0
3      0    5 -0.47 36.32        0        6    0      0
4      1    6 -0.37 34.89        0        8    0      0
5      1    0 -0.53 43.58        0        4    1      0
6      0    0 -0.48 41.25        0        4    1      0
  service sports music student_gov honor journalism vocation
1       0      0     0           0     0          0        0
2       0      0     0           0     0          0        0
3       0      0     0           0     0          0        0
4       0      0     0           0     0          0        0
5       0      0     0           0     0          0        0
6       0      0     0           0     0          0        0
  income2011
1          0
2          0
3          0
4          0
5          0
6          0

> attach(els2002) #allows R to recognize variable names
```

Now that you've input the data set we'll work with throughout the e-book, you're all set up to begin learning and performing the analyses. The first three chapters will describe prerequisites to any analysis: describing and visualizing data. Let's get started.

Chapter 1 Central Tendency

Calculate measures of center

Central tendency is a term that describes one point at which a group of values gathers. **Measures of center** are statistics that describe the central tendency. You've probably heard of the three most commonly used measures of center: **mean**, **median**, and **mode**. This chapter will define and describe how to calculate each. It is important to know what they are, and it is also important to know what they mean, especially in relation to each other. By looking at how the mean, median, and mode compare to each other, we can better understand the story the data is telling.

Mode

The mode is where most of the numbers in a data set occur, i.e. where the **frequency** is the greatest. This could be a single number or a group of numbers. For example, in the small data set {2, 3, 3, 4, 6}, the mode is 3 because the frequency is 2 (3 appears twice), while the frequency of the other values is 1.

A **histogram** is the most common way to visualize the mode of a distribution. A histogram is a special type of bar graph that shows the values in the data set on the x-axis and the frequency of those values on the y-axis. Values on the x-axis are grouped into bins of a specified range or category.

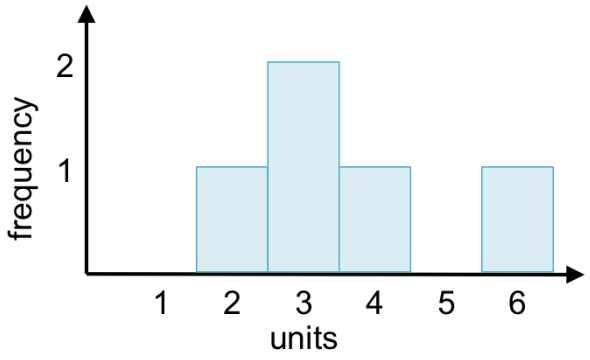


Figure 1: A histogram of the data set {2, 3, 3, 4, 6} shows that the mode is 3.

The mode doesn't have to be a number. For example, let's say in the local high school biology class, 10% of students scored A's, 40% scored B's, 35% scored C's, and 15% failed. In this case, the mode is a grade of B. This type of data set is an example of **categorical data** (as opposed to **numerical data**), in which data is arranged in categories (in this case, the grade in the class).

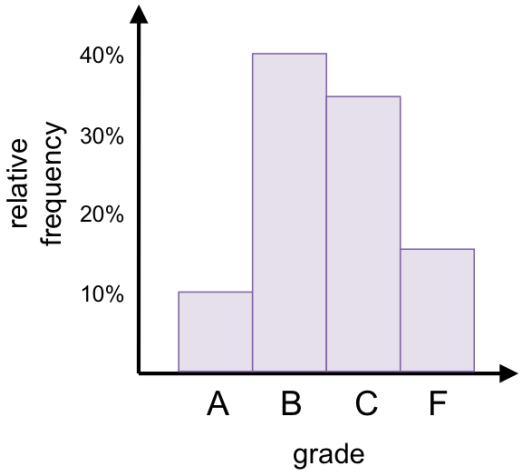


Figure 2: In this categorical data set, the mode is a grade of B.

Note: The y-axis in Figure 2 shows the relative frequency—the frequency of each category in relation to one another—rather than the absolute frequency, which would depict the absolute number in each category.

For large, continuous data sets, the mode is the range with the highest frequency. The following histogram shows ten bins, each with a width of 5 units. You can easily see that the mode is the range (35, 40). Remember that the mode is *where* the greatest frequency occurs (the values along the x-axis), not *what* the frequency is (i.e. the mode is not 8).

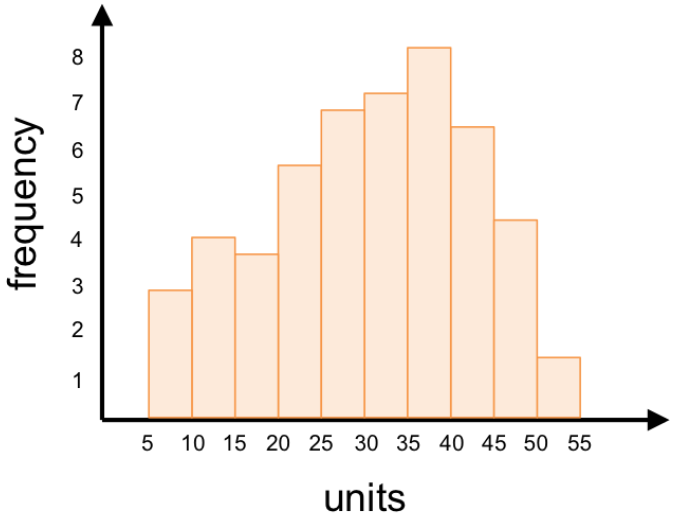


Figure 3: The mode of the data set visualized by this histogram is the range (35, 40).

Median

The median is another measure of center. This statistic is a number for which 50% of the values in the data set are less and 50% are greater. In the case of a data set with an odd number of values, the median is an actual value in the data set and is smack dab in the middle. For example, the data set {5, 6, 8, 12, 15} has a median of 8. Two values are less than 8 and two values are greater than 8.

When a data set has an even number of values, the median is the average of the two middle numbers. The median of the data set {4, 6, 9, 11, 17, 18} is 10—the average of 9 and 11. Three values are less than 10 and three values are greater than 10.

You may have noticed that values must be ordered; otherwise the median can't be computed. Also note that we can't find the median for categorical data, but we can for numerical data.

Note: Unlike the mode, we can't easily see where the median is by looking at a histogram. We need to put the values in order and find the middle number(s).

Notice that the numbers less than or greater than the median can be anything (so long as they remain less than or greater than the median) and the median will remain the same. For example, the following data sets all have the same median:

{5, 6, 8, 12, 15}

{5, 6, 8, 20, 300}

{-100, -16, 8, 12, 15}

{-100, -16, 8, 20, 300}

Therefore, the median by itself does not adequately describe a data set.

It's also helpful to have a statistic that accounts for every value. This is why a more common measure of center is the mean.

Mean

Unlike the mode and median, the mean (also known as the arithmetic mean) uses every value in the data set in its calculation.

Note: Other types of means exist (e.g., geometric mean, harmonic mean), but the arithmetic mean is the most common. It is "arithmetic" because it's calculated by* adding *every value in the data set and then dividing by the number of values.

For a data set $\{x_1, x_2, x_3, \ldots x_n\}$, where n is the number of values in the data set, the mean is represented by $\bar{x}$ (x-bar) and is equal to:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

We can rewrite this as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The Greek letter capital sigma (Σ) symbolizes taking the sum. The $i = 1$ and the n on the bottom and top of Σ indicate the values of i that we should use in the summation: 1, 2, 3, … n. So, we should substitute the subscript i from x_i with 1, 2, 3, all the way to n (where n could be any number). This translates to finding the sum of x_1, x_2, x_3, all the way to x_n. Then we divide the sum by n to find the mean.

We use the symbol $\bar{x}$ (x-bar) to represent the mean of a sample and the symbol μ (mu) to represent the mean of a population. In general, we use lowercase letters when describing a sample and uppercase letters when describing a population (x_i are values of a sample and n is the sample size, while X_i are values of a population and N is the population size). Therefore, the mean of a population is represented by:

$$\mu = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Note: If you're wondering why statistical notation must be so complicated, you're not alone. But once you get the hang of it, this language is a very useful tool for quickly and easily communicating complex statistical ideas.

Because the mean uses every value in its calculation, **outliers** (values in the data set that differ significantly from other values in the same data set) can severely affect it. Take the two data sets below, one of which has an outlier:

{4, 6, 7, 10} $\bar{x}$ = 6.75

{4, 6, 7, 100} $\bar{x}$ = 29.25

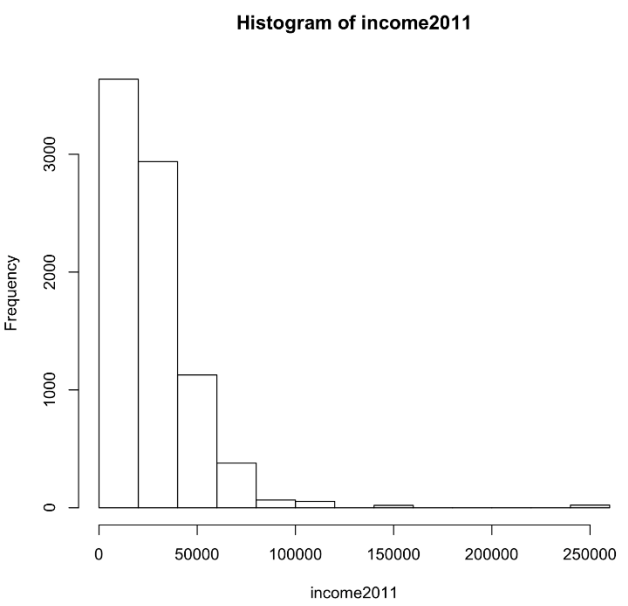
This example illustrates why the mean is not always the best measure of center. If we only know the mean of the second data set, we would think that the values cluster around 29.25, when in fact 75% of them are less than 8.

When the mean, median, and mode for a data set are roughly equal, the mean is used to calculate many other statistics (e.g., how spread out the data is) to perform a multitude of analyses. For this reason, the mean is the most common measure of center.

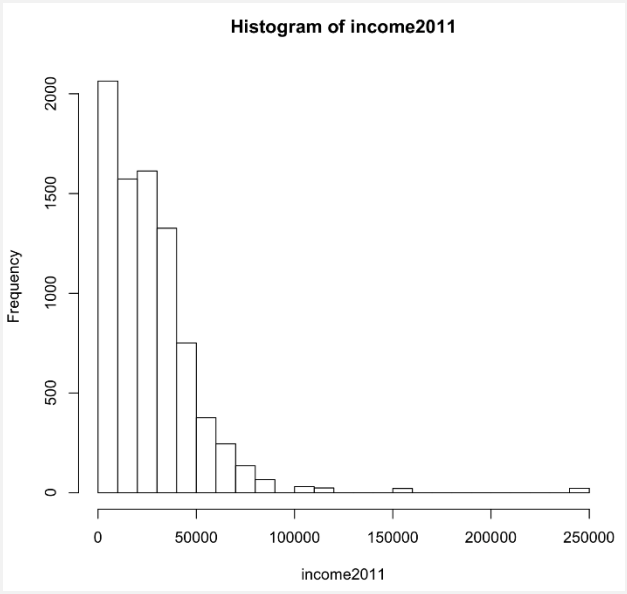
Let's look at how to find the mean, median, and mode in R using the NCES data. (If you have not yet downloaded the data, imported it into R, and run the **`attach()`** function, do this first.)

Code Listing 2

```
> mean(income2011) #outputs the mean of respondents' income in 2011
[1] 27302
> median(income2011) #outputs the median of respondents' income in 2011
[1] 24000
> hist(income2011) #outputs a histogram of respondents' income in 2011
```




```
> hist(income2011, breaks=20) #outputs a histogram with smaller bin sizes
```



You can see that the variable “income2011” is heavily skewed (i.e. most values fall on one side of the full range of the data). The majority of students have a family income of less than $50,000.

This skewedness results in the mean being greater than the median. Recall that the median is not influenced by outliers because it is the exact middle value, while the mean is affected by every value in the data set. In this case, outliers (students with a family income of $250,000) are pulling the mean to the right. When the mean differs from the median, it suggests the presence of outliers and a skewed distribution such as the one in this example.

Taken together, the mean, median, and mode can provide a useful description of a data set. In the next chapter, you’ll learn methods to calculate the variability of a data set—in other words, how “spread out” values are in relation to each other. When describing data, the variability is as important as the measures of center.

Chapter 2 Variability

Calculate measures of spread

While measures of center describe where values in a data set gather, **measures of spread** describe a data set's **variability** (how spread out values are).

Consider this simple example that illustrates why we might want to know variability in the real world—the task of deciding which clothes to bring on your vacation. Let's say the average temperature in your destination city is 74°F. Based on this number alone, you would probably bring only shorts and T-shirts. But what if the high is 102°F and the low is 34°F? After knowing this range, you would probably want to bring a coat and a bathing suit as well.

Throughout this chapter you'll learn common methods to measure spread.

Range

In the temperature example, the **range** is a useful measure of spread. To calculate it, subtract the minimum value from the maximum value: 102°F – 34°F = 68°F. That's a pretty big difference in temperature! Range is the simplest measure of spread. It can tell part of the story, but, like many statistics in isolation, range can sometimes be deceiving.

For example, let's say you want to buy a house. You analyze the property values of other houses in the area and find that in one particular neighborhood, the houses have the following values:

$355,000
$299,995
$323,500
$286,350
$333,290
$410,280
$810,975

The range is pretty large ($810,975-$286,350 = $324,625). You can see from looking at the data that one of the houses has a much larger property value than the others ($810,975) and that the other property values are in the $200K-$400K range. So, rather than using a simple range as a measure of spread, you might consider the **interquartile range (IQR)**.

IQR

We know now that the median splits the data in half. Using this same method, we can split the data into fourths, so that 25% of the data is less than the first value, 25% is between the first and second value, 25% is between the second and third value, and 25% is greater than the third value. In doing this, we will find three values we'll use to calculate the IQR.

The first value is called the first quartile, abbreviated Q_1; the second value is the median, also known as the second quartile and abbreviated Q_2; and the third value is the third quartile, abbreviated Q_3. The difference between Q_3 and Q_1 (in other words, the middle 50% of the data set) is the IQR, and statisticians often calculate this in order to reduce the impact that outliers can have on the range calculation.

To make this calculation, we need to place the values in numerical order.

After ordering the data, we first find the median. Then, we find the median of each half of the data (Q1 and Q3). Note that the calculation of Q1 and Q3 do not include the median value.

Table 1: The data set we'll use to find the IRQ for a range of neighborhood housing prices.

Quartile	House Value
	$286,350
Q_1	**$299,995**
	$323,500
Q_2	**$333,290**
	$410,280
Q_3	**$355,000**
	$810,975

In this example, IQR = Q_3–Q_1 = $355,000 - $299,995 = $55,005.

Outliers are formally defined as any values that are either less than Q_1 - 1.5(IQR) or greater than Q_3 + 1.5(IQR). In this case, a value is an outlier if it is less than $299,995 – 1.5($55,005) = $217,487.5 or greater than $355,000 + 1.5($55,005) = $437,507.5. In this case, the only outlier is $810,975. Outliers are represented by dots on a box plot.

A **box plot** visualizes where the minimum and maximum values, Q_1, Q_2, Q_3, and any outliers are in relation to each other. Note that the minimum and maximum values are the smallest and largest values that are not considered outliers.

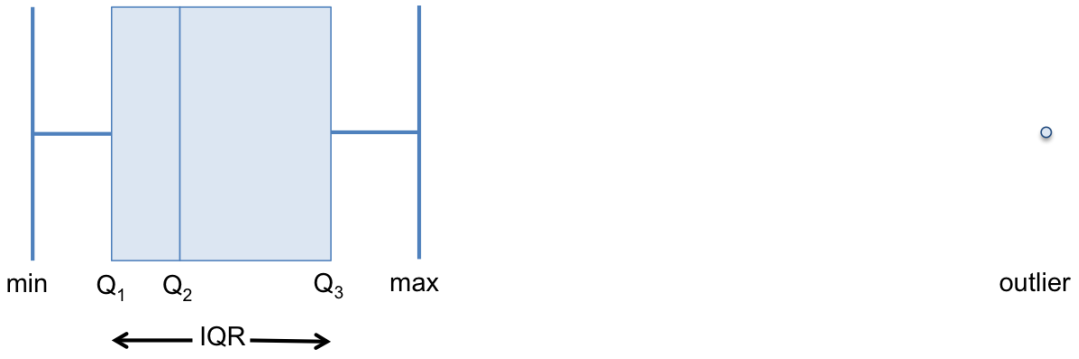


Figure 4: Box plots visualize where the minimum value, first quartile (cutoff of smallest 25% of values), second quartile (i.e. median), third quartile (cutoff of largest 25% of values), maximum value, and outliers are in relation to each other.

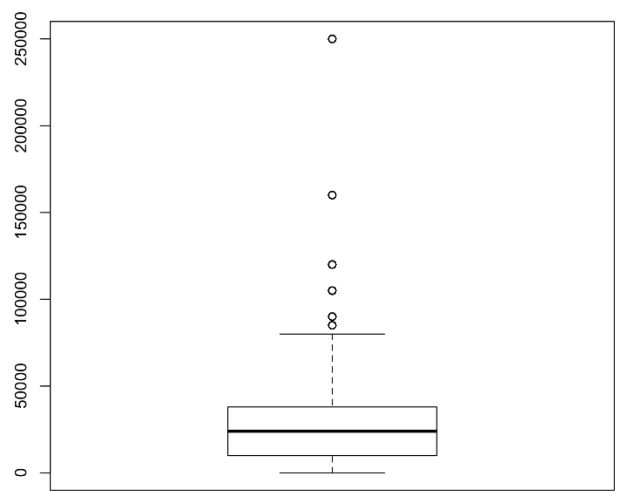
You'll now use R to find the minimum value, first quartile, second quartile, third quartile, and maximum value in order to graph a box plot.

Code Listing 3

```
> summary(income2011) #outputs the min, Q1, Q2, Q3, max, and mean

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   10000   24000   27300   38000  250000

> boxplot(income2011) #creates a box plot
```




```
> IQR(income2011) #calculates the IQR

[1] 28000
```

The IQR can be a useful statistic for spread, but notice that the IQR only takes two values (Q_1 and Q_3) into account. In other words, any of the other values can change (as long as they remain between or outside Q_1 and Q_3—whatever they were originally) and the IQR will stay the same.

Therefore, we tend to use the **standard deviation,** which takes every value in the data set into account, more commonly than the IQR.

Standard deviation

Before learning what the standard deviation is or how it's calculated, let's first consider how we might use every value in a data set to compute a single statistic that measures spread.

Consider this sample data set: {11, 10, 4, 12, 15, 8, 14, 6}.

Now look at each individual value's **deviation** from the mean (i.e. the distance between each value and the mean, equal to $x_i - \bar{x}$). For this data set, the mean ($\bar{x}$) is 10.

Table 2: Finding each individual value's deviation from a mean of 10.

x_i	$x_i - \bar{x}$
11	1
10	0
4	-6
12	2
15	5
8	-2
14	4
6	-4

You can calculate the average deviation to find the average difference that a value lies from the mean. Makes sense, right? However, if you calculate the average deviation, you get 0. You can see that algebraically the sum of the deviations is equal to 0 (and therefore the average is 0 as well):

$$\Sigma(x_i - \bar{x}) = \Sigma\left(x_i - \frac{\Sigma x_i}{n}\right) = \Sigma\left(\frac{nx_i}{n} - \frac{\Sigma x_i}{n}\right) = \Sigma\frac{1}{n}(nx_i - \Sigma x_i) = \frac{1}{n}\Sigma(nx_i - \Sigma x_i)$$

And we know that $\Sigma(nx_i - \Sigma x_i) = 0$ because

$$\Sigma(nx_i - \Sigma x_i) = (nx_1 - \Sigma x_i) + (nx_2 - \Sigma x_i) + \cdots + (nx_n - \Sigma x_i)$$

$$= nx_1 + nx_2 + \cdots + nx_n - n\Sigma x_i$$

$$= n(x_1 + x_2 + \cdots + x_n) - n\Sigma x_i$$

$$= n\Sigma x_i - n\Sigma x_i$$

$$= 0$$

The average deviation is equal to 0, so it isn't much help as a statistic for spread. To solve this problem, you could use the average absolute deviation, where each absolute deviation is $|x_i - \bar{x}|$. (The notation || takes the absolute value of a number.)

Table 3: The average absolute deviation is found using the data set: {11, 10, 4, 12, 15, 8, 14, 6} and each absolute deviation.

x_i	$\lvert x_i - \bar{x}\rvert$
11	1
10	0
4	6
12	2
15	5
8	2
14	4
6	4

If you take the average absolute deviation, you get 3. So the average distance of each value from the mean is 3.

The average absolute deviation works as a measure of spread, but the **standard deviation** is used more commonly. Instead of taking the absolute value of each deviation, with the standard deviation we square each deviation (remember that squaring a number will always produce a positive value), find the average squared deviation, and then take the square root. If we look at just one of the values in the data set (6), we can visualize the deviation and squared deviation.

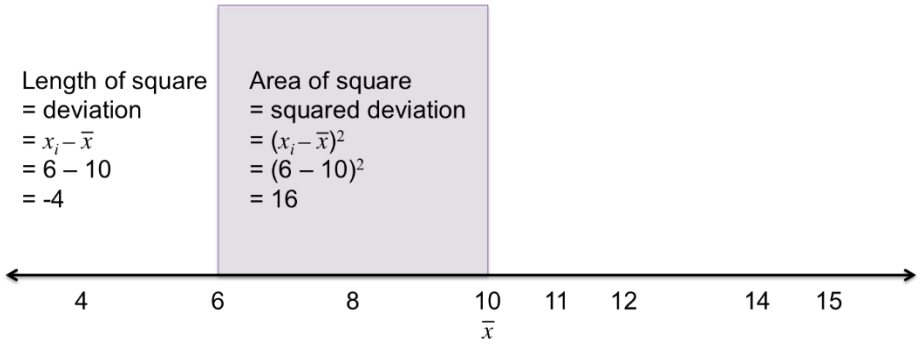


Figure 5: This figure visualizes the deviation $x_i - \bar{x}$ *by the side length of the square, and the squared deviation* $(x_i - \bar{x})^2$ *by the area of the square.*

If we calculate each squared deviation and take the average, we get a measure of spread called the **variance (σ^2)**. In our example, the variance is 12.75.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

The numerator of the expression for variance is often referred to as the **sum-of-squares (SS)**. This should make sense, as this is the sum of each squared deviation.

If we take the square root of the variance, we get the standard deviation (σ):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Essentially, this is finding the side length of the average squared deviation.

Using Figure 6, we can visualize each value in the data set (4, 6, 8, 10, 11, 12, 14, 15), the mean (10), each squared deviation (purple squares), the variance (orange square), and the standard deviation (side length of the orange square).

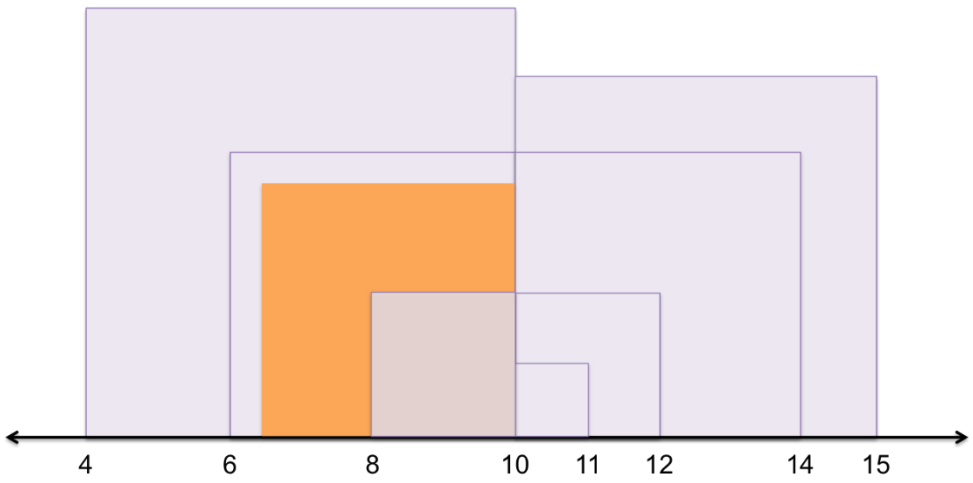


Figure 6: This figure visualizes each deviation from the mean (10) by the side lengths of each square; each squared deviation by the area of each square; and the average squared deviation by the orange square.

Let's calculate the standard deviation in our example.

Table 4: The standard deviation is found using the data set, the absolute deviation, and the square of each absolute deviation.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
11	1	1
10	0	0
4	6	36
12	2	4
15	5	25
8	2	4
14	4	16
6	4	16

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{1^2 + 0^2 + 36^2 + 4^2 + 25^2 + 4^2 + 16^2 + 16^2}{8} = 12.75$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{12.75} = 3.57$$

So, here the standard distance between each value and the mean is 3.57.

This calculation for standard deviation is used for a population (i.e. when we have all values of a certain variable). However, often we don't have data for the entire population, so we have to use a sample (a smaller subset) to draw conclusions. Frequently, the sample will have a smaller variance than the population because randomly chosen values are likely to be closer to the measures of center. Therefore, to better approximate σ (the standard deviation of the population), we subtract 1 from the denominator to make the whole calculation slightly larger. We denote this approximation s, and refer to it as the **sample standard deviation**.

$$\sigma \approx s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

By default, R calculates the sample standard deviation. However, you can calculate the population standard deviation with simple algebra.

Code Listing 4

```
> sqrt(sum((income2011-mean(income2011))^2)/length(income2011))
  #outputs the population standard deviation (the sum-of-squares divided
  by n)

[1] 24531.3

> sd(income2011) #outputs the sample standard deviation (s)

[1] 24532.78
```

The standard deviation and sample standard deviation are used for a variety of statistical tests that enable you to draw conclusions and make decisions based on what the data tells you. You'll begin learning these tests in Chapter 4, but for now you should know that the mean and the standard deviation can help you determine if a value is likely or unlikely to occur. If you know that most values in a data set are a certain distance from the mean, and you get a value that is a lot farther from the mean, you know something weird is going on.

Before getting into statistical testing, we'll look into the shape of distributions—one more factor to consider when getting to know your data.

Chapter 3 Distributions

Visualize the shape of data

Measures of center and spread tell part of the story. We also need to look at the shape of the **distribution** by creating histograms. Histograms are a special type of bar graph that shows the frequency of values in a data set that lie between evenly spaced intervals. A distribution, on the other hand, is a theoretical curve that models a histogram's shape. We can use these curves to calculate estimated probabilities on which to base our conclusions.

We'll now look at examples of normal, uniform, skewed, and bimodal distributions. One of the most important is the normal distribution.

Normal distribution

In a perfectly normal distribution, the mean, median, and mode are equal, and they are exactly in the middle of the data set, with frequencies **symmetrical** (the same number of values occurs on either side of the median). While real-life data sets are never perfectly normally distributed, we can model them with a theoretical normal curve.

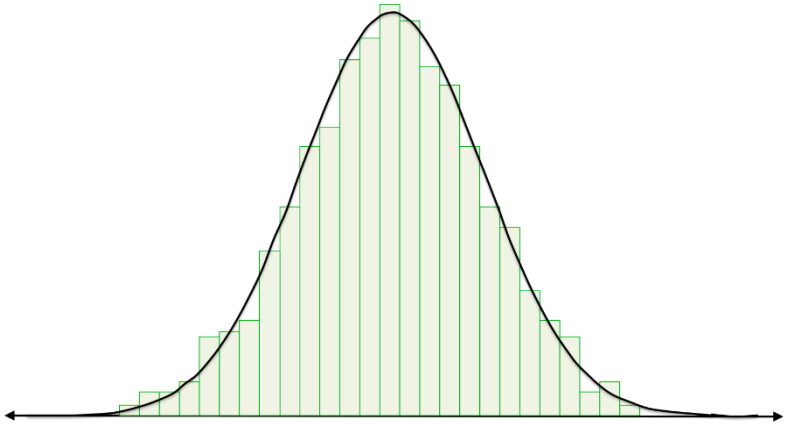


Figure 7: The normal distribution (depicted by the black curve) can be used to model relatively normal data sets (visualized by the green histogram).

This curve has the equation

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean and σ is the standard deviation.

The area under the normal curve is equal to the probability; thus, the total area under the curve is equal to 1. This should make sense intuitively if you look at the histogram. Using Figure 7 as an example, what is the probability that if you randomly select a value in the data set, it will be in one of the bins depicted by the green bars? The probability is 100%.

Now a slightly more complex question: How would you calculate the probability of randomly selecting a value from the data set that is in one of the five leftmost bins? You would sum the frequency of values in each of those bins and divide by the total frequency. Do you see now why the area under the normal curve is the probability? Therefore, the equation for the curve that models a data set's distribution is called the **probability density function (PDF)**.

A perfectly normal distribution with $\mu = 6$ and $\sigma = 2.1$ has the following equation for its PDF:

$$N(6, 2.1) = \frac{1}{\sqrt{2\pi(2.1)^2}} e^{-\frac{(x-6)^2}{2(2.1)^2}} = 0.19e^{-\frac{(x-6)^2}{8.82}}$$

If we graph this, we get Figure 8.

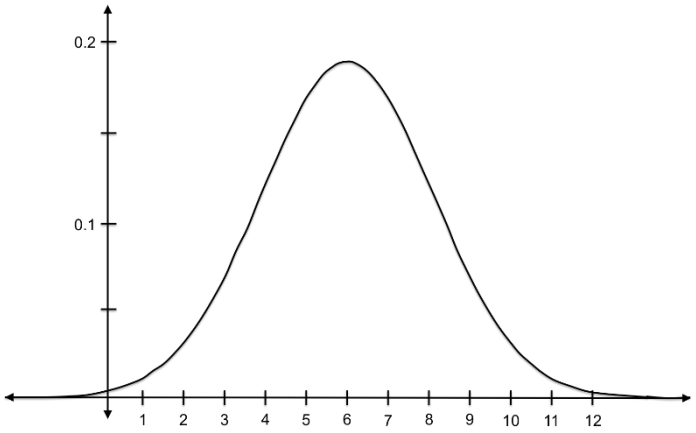


Figure 8: This curve is the PDF for a normal distribution with a mean of 6 and a standard deviation of 2.1.

Let's look at the probability of randomly choosing a value that is less than 4:

$$P(x < 4) = \int_{-\infty}^{4} N(6, 2.1) \approx 17\%$$

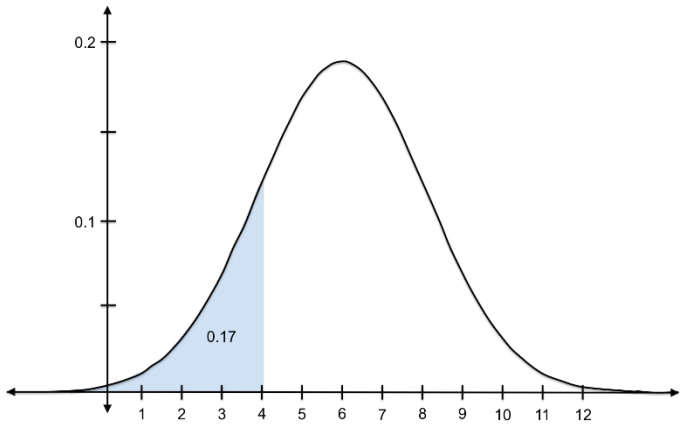


Figure 9: The probability of randomly selecting a value less than 4 from a distribution with $\bar{x}$ *= 6 and* σ *= 2.1 is 0.17.*

Here is the probability of randomly choosing a value between 5 and 8:

$$P(5 < x < 8) = \int_5^8 N(6, 2.1) \approx 51\%$$

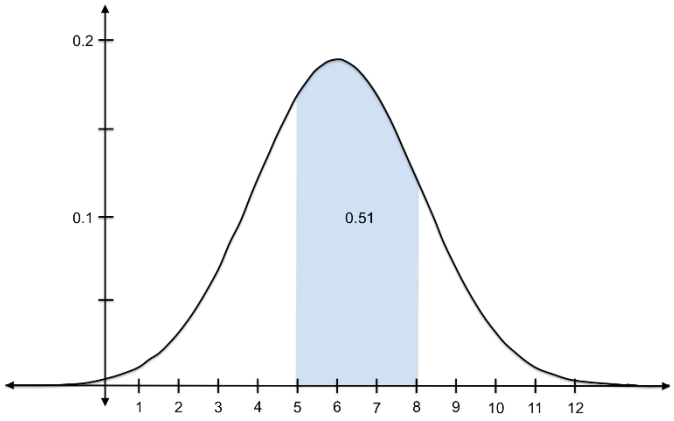


Figure 10: The probability of randomly selecting a value between 5 and 8 from a distribution with $\bar{x}$ *= 6 and* σ *= 2.1 is 0.51.*

We find these probabilities by integrating the PDFs. (For anyone who doesn't remember calculus all too well, taking the integral is the opposite of taking the derivative. Integrating gives us a function that, when plugged into values of x, results in the cumulative area under the original curve up to x.) However, we don't need to continue integrating each PDF in order to calculate probabilities. To make things easier on ourselves, we can **standardize** the normal distributions by converting each into a **standard normal distribution** with mean 0 and a standard deviation 1. This special normal distribution is denoted *N*(0,1). If we replace μ and σ, we get:

$$N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

We then use a special table that gives us cumulative probabilities under the standard normal curve. You will learn how to standardize normal distributions and calculate probabilities in Chapter 4.

In this e-book, we'll focus on normal distributions, but first let's explore some other common distributions.

Uniform distribution

Data is uniformly distributed when the probability of randomly selecting a particular value is about the same as that for another value. In other words, when the frequency in each bin is the same. Therefore, the theoretical uniform distribution is perfectly horizontal.

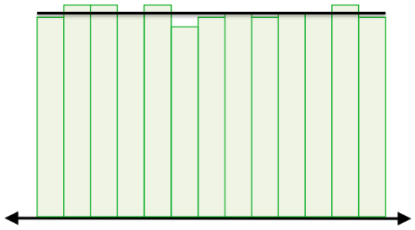


Figure 11: This is a uniform distribution because the frequency in each bin remains relatively constant.

The equation is written as $U(a,b) = \frac{1}{b-a}$, where a is the minimum value and b is the maximum value.

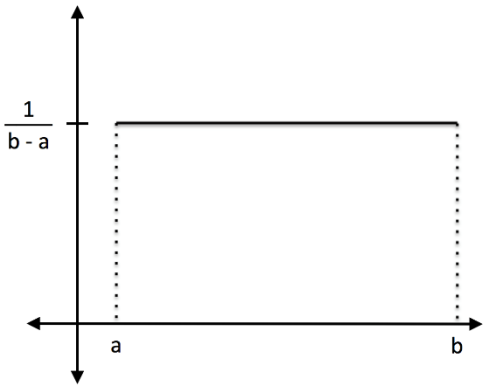


Figure 12: This line is the PDF for a uniform distribution with minimum a and maximum b.

You can see from looking at the uniform distribution that the area under the curve is 1. The probability of randomly selecting a value less than q is:

$$P(x < q) = \int_a^q \frac{1}{b-a} = \frac{q-a}{b-a}$$

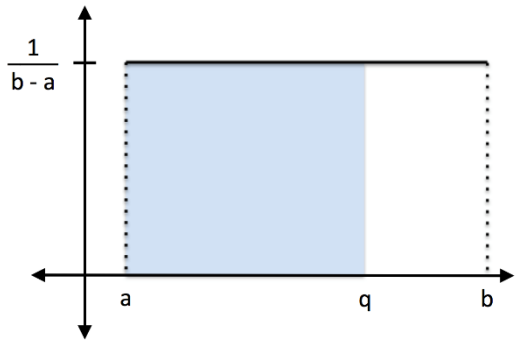


Figure 13: The probability of randomly selecting a value less than q from a uniform distribution with minimum a and maximum b is (q-a)/(b-a).

The outcomes of rolling a die offer a good example of a uniform distribution. Each number has an equal probability of being selected, so if you rolled 600 times, you should get around 100 of each value (unless of course, the die was rigged).

Skewed and bimodal distributions

Skewed and bimodal distributions are also very common. Income is one example of a heavily skewed distribution—the wealthiest 10% of Americans own 75% of all wealth in America.[1]

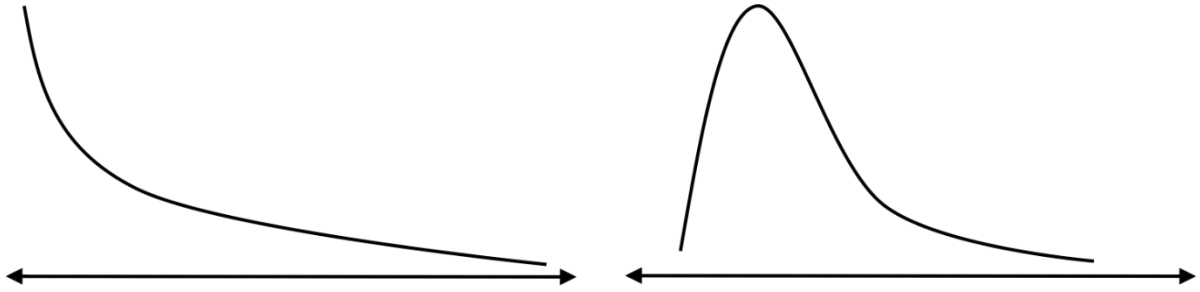


Figure 14: Skewed distributions have the highest frequencies occurring on one end of the range.

[1] Zuesse, E. (2015). *U.S. Wealth-Concentration: Wealthiest Tenth (10%) of Americans Own 75% of America.* Center for Research on Globalization.

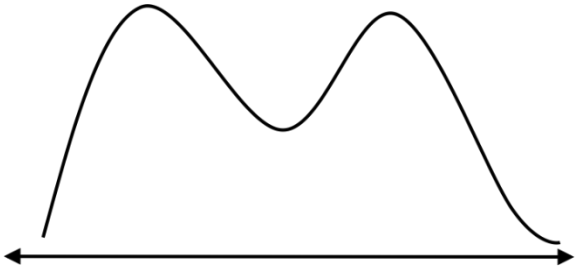


Figure 15: Bimodal distributions have the highest frequencies occurring in two areas of the range.

Visualizing data and calculating descriptive statistics (measures of center and spread) are important precursors to any analysis. Most of the statistical analyses presented here is used when we have normally distributed data, but in Chapters 5 and 6 you'll also learn how to draw conclusions about samples drawn from data of any distribution.

Chapter 4 Standardizing

Use distributions to find probabilities

In the previous chapter, we looked at converting normal distributions into the standard normal distribution ($\mu = 0$, $\sigma = 1$) so that we don't have to integrate the PDF to calculate probabilities. This process of converting any normal distribution into a standard normal distribution is called **standardizing**. It allows us to compare two different normal distributions. Specifically, say you are analyzing a value from one normal distribution along with another value on a different normal distribution. How would you know which value is farther from the norm, based on the distribution it comes from? That's what you'll learn in this chapter.

Consider the following distribution with $\mu = 15$ and $\sigma = 3$.

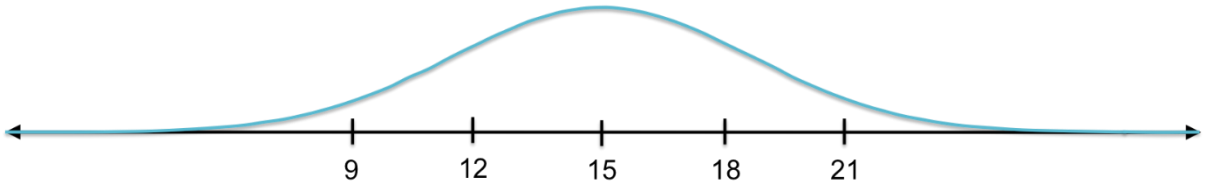


Figure 16: This curve is the PDF for a normal distribution with a mean of 15 and a standard deviation of 3.

How would we convert this into a standard normal distribution (mean = 0 and standard deviation = 1)?

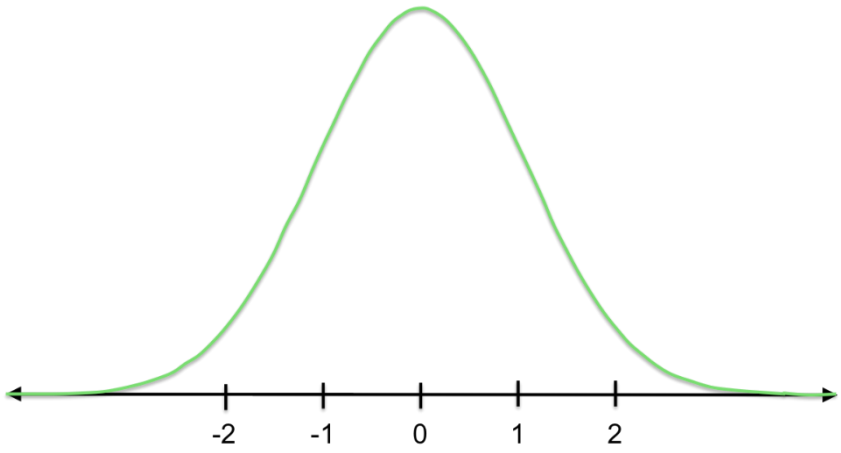


Figure 17: This curve is a standard normal distribution N(0,1), which has mean 0 and standard deviation 1. When we standardize normal distributions, we convert them into the standard normal distribution.

To put the question another way, let's say that 21 is one of the values in the original data set. If we shift and shrink the distribution to become standard normal, what new value will 21 have?

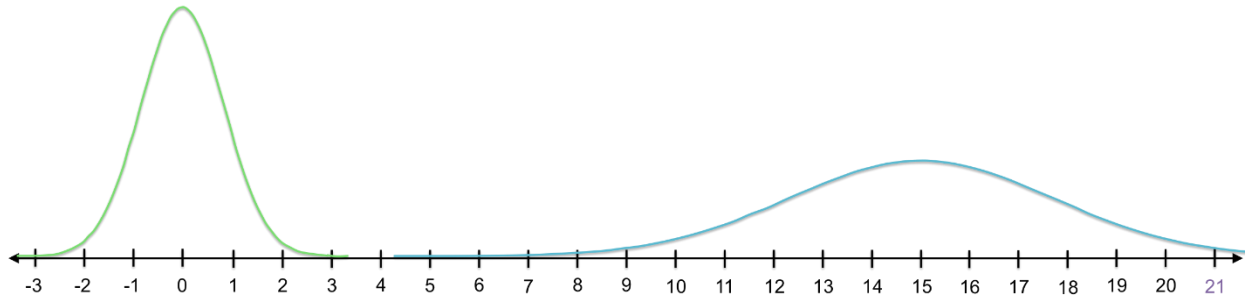


Figure 18: If we want to standardize the blue curve, we need to develop a system for converting it into the green (standard normal) curve.

You might have guessed that we first need to subtract the mean from 21. More broadly, in order to convert the entire distribution into the standard normal distribution, we first must subtract the mean from each value in the data set (x_i - μ). This shifts the data to the left so that the new mean is 0.

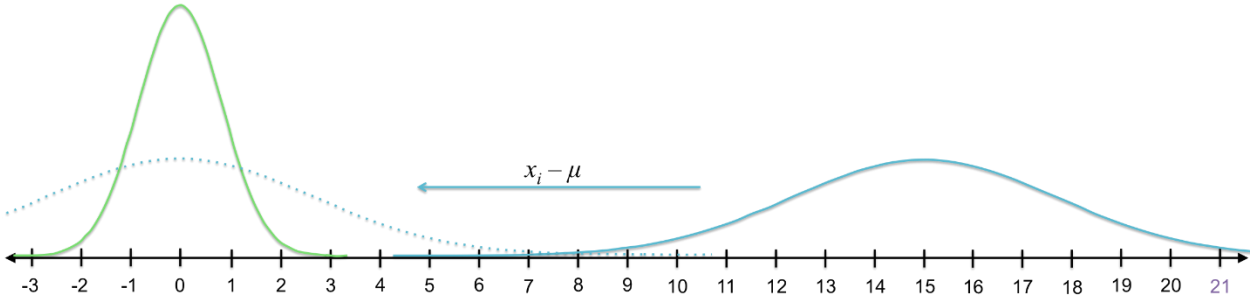


Figure 19: To convert the blue curve on the right into the green (standard normal) curve on the left, we subtract the mean of the blue distribution from each value in the distribution (shifting the entire distribution to have a mean of 0) and then divide by the standard deviation (shrinking the distribution to match the green distribution).

You can see that the new value for 21 would be 6.

Now the mean of each distribution is the same, but the standard deviation of our original data set is 3 and we want it to be 1. So, how can we shrink the spread? We divide by the standard deviation. The standardized value of 21 is therefore (21-15)/3 = 2.

In our original data set, 21 is two standard deviations from the mean (recall the standard deviation is 3 and the mean is 15). When we standardize 21 and it becomes 2, it remains two standard deviations from the mean (now the mean is 0 and the standard deviation is 1).

Think of standardizing this way—you're simply calculating the number of standard deviations a value is from the mean. This number is called the **z-score**, denoted z.

$$z = \frac{x_i - \mu}{\sigma}$$

Values from different normally distributed data sets with the same z-score will have the same probability of being selected. In other words, the area under the curve that is greater than or less than values with the same z-scores is the same. This is because the total probability under any PDF is 1.

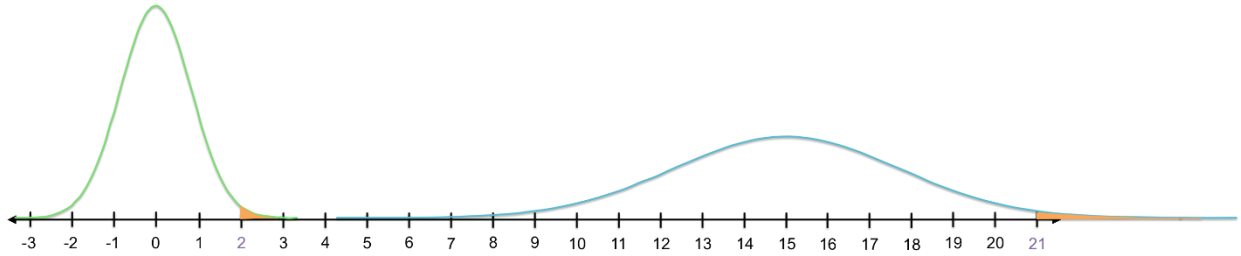


Figure 20: The area under the standard normal curve above 2 is equal to the area under the original curve (μ = 15, σ = 3) above 21. Both 2 and 21 are two standard deviations above the mean of their respective data sets.

By standardizing distributions and finding z-scores for each value of interest, we can use the standard normal distribution to calculate all our probabilities (i.e. the areas under the PDF). The z-table located at the end of this e-book lists cumulative probabilities for any z-score.

The numbers in the body of that table are the cumulative probabilities (p) less than a particular z-score. For example, the probability of randomly choosing a value less than z = 1.22 is 0.8888 (i.e. P(x < 1.22) = 0.8888).

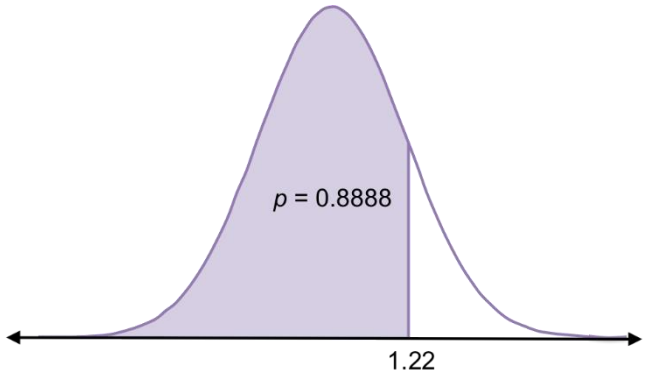


Figure 21: In the standard normal curve, the probability of randomly selecting a value less than 1.22 (i.e. 1.22 standard deviations above the mean) is 0.89.

Note: In probability notation, we usually use x to denote the value of the variable. Note that in this example, because our values of interest follow a standard normal distribution, the x-values are also z-scores.

Upon looking at the z-score table, you'll see that it shows only cumulative probabilities for positive z-scores. However, since the normal distribution is symmetrical, you can also use this table to calculate probabilities for negative z-scores. For example:

P(x > -1.22) = 0.8888
P(x < -1.22) = 1 – 0.8888 = 0.1112

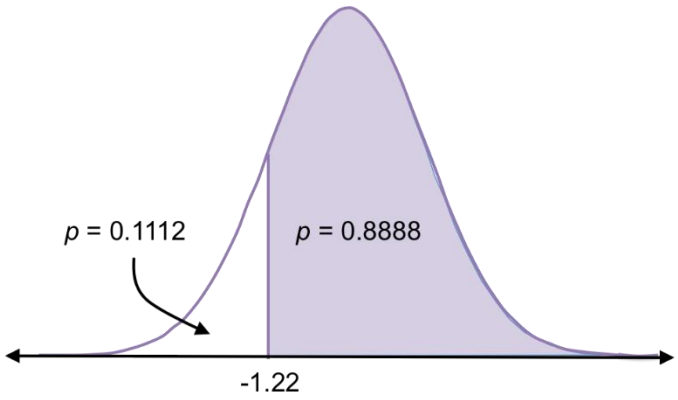


Figure 22: Because normal distributions are symmetric, the probability of randomly selecting a value less than +1.22 is the same as randomly selecting a value greater than -1.22, which is 0.89. All probabilities under the curve add to 1, which means the probability of randomly selecting a value greater than +1.22 is 1 – 0.89 = 0.11, and this is the same as the probability of selecting a value less than -1.22.

Now you know how to do basic statistical analyses. Given any data set, you can describe it, visualize it, and calculate the probability of a particular range of values occurring (using the z-table for normally distributed data or using the PDFs to model data of other distributions).

Let's now work through a simple real-world example.

Example

You want to take guitar lessons. Somehow you know that the hourly rate for guitar teachers in your area is normally distributed with μ = \$28 and σ = \$4.6. If you look at a list of guitar teachers' contact info and randomly choose one to call, what is the probability that this teacher charges between \$15 (the minimum for getting a decent teacher) and \$25 (the maximum you're willing to pay)?

First, we'll standardize the distribution by finding the z-scores for \$15 and \$25. Because this price range is below the mean, we will expect negative z-scores.

$$z = \frac{15 - 28}{4.6} = -2.83$$

$$z = \frac{25 - 28}{4.6} = -0.65$$

So, you want to find the area under the standard normal curve between -2.83 and -0.65. In other words, you're looking for the probability of randomly selecting a value between -2.83 standard deviations from the mean and -0.65 standard deviations from the mean.

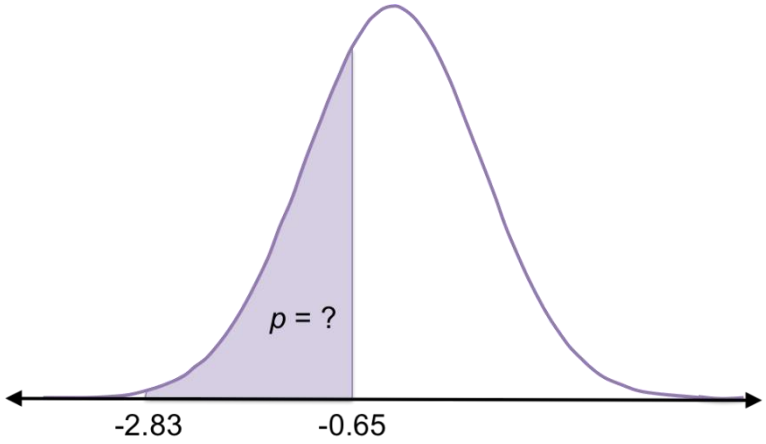


Figure 23: You can use the z-table to find the probability of randomly selecting a value between -2.83 and -0.65 standard deviations from the mean for a normal distribution.

Now you can use the z-table to find the cumulative probability less than positive 2.83, then subtract the cumulative probability less than positive 0.65. Because the normal distribution is symmetric, this is the same probability depicted in Figure 23.

$P(x < 2.83) = P(x > -2.83) = 0.9977$
$P(x < 0.65) = P(x > -0.65) = 0.7422$
$P(0.65 < x < 2.83) = P(-2.83 < x < -0.65) = 0.9977 – 0.7422 = 0.2555$

Therefore, the probability of randomly selecting a guitar teacher who charges between $15 and $25 per hour is about 0.2555, or 25.55%. That means you'd likely find a guitar teacher after four phone calls.

Determine what is significantly unlikely

The shape of the normal distribution—high frequencies around the mean, median, and mode and low frequencies in the tails—allows us to determine if something weird is going on with a particular value or sample (i.e. if we randomly selected a value or sample that is extremely unlikely to be randomly selected).

There are many situations in which we might want to statistically determine if a value is significantly different from the mean. One area is in health: for example, knowing if your heart rate or cholesterol levels are unhealthily high or low.

In normal distributions, a majority of values (about 68%) lie within 1 standard deviation of the mean and almost all (about 95%) lie within 2 standard deviations of the mean.

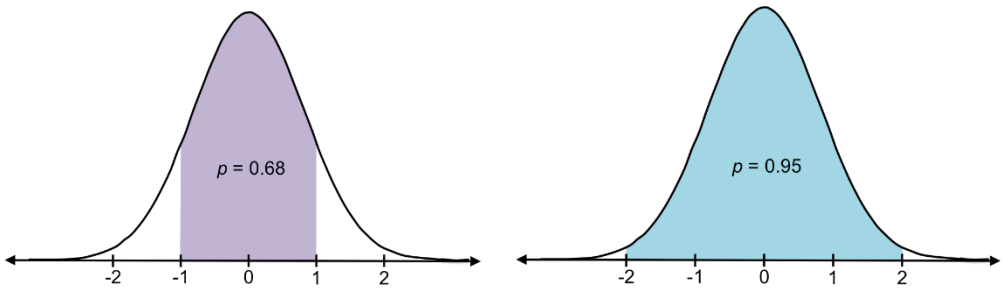


Figure 24: In normal distributions, about 68% of values lie within one standard deviation of the mean, and about 95% of values lie within two standard deviations of the mean.

Therefore, randomly selecting a value more than two standard deviations from the mean in either direction is very unlikely. Generally, we decide that something is statistically unlikely if the probability of selecting a value is less than 0.05. It's even more unlikely to occur if the probability is less than 0.01, and really *really* unlikely if the probability is 0.001. These probabilities (0.05, 0.01, and 0.001) are known as **alpha levels (α)**, also called significance levels because if the probability of selecting a value or sample is less than α, the results are considered "significant."

For example, in the guitar-lesson example, the z-score for $15 is -2.83. This is more than two standard deviations below the mean, i.e. finding a guitar teacher who charges $15 per hour or less is statistically unlikely.

Determining whether or not a probability is less than α is called **hypothesis testing**. This chapter covers **z-tests**: hypothesis testing when we know population parameters μ and σ. For this test, we continue using the z-table. (When we don't know population parameters, we have a different distribution and use a different table. This will be the focus of Chapter 6.)

We can use three types of hypothesis tests:

- Left-tailed test
- Right-tailed test
- Two-tailed test

All tests use the same alpha levels; however, each has a different location for the cutoff between what is considered significant or not. A left-tailed test analyzes whether or not a value or sample falls significantly below the mean (i.e. in the bottom α); a right-tailed test analyzes whether or not a value or sample falls significantly above the mean (i.e. in the top α); and a two-tailed test analyzes whether or not a value or sample is significantly different from the mean in either direction (i.e. in the bottom $\alpha/2$ or top $\alpha/2$).

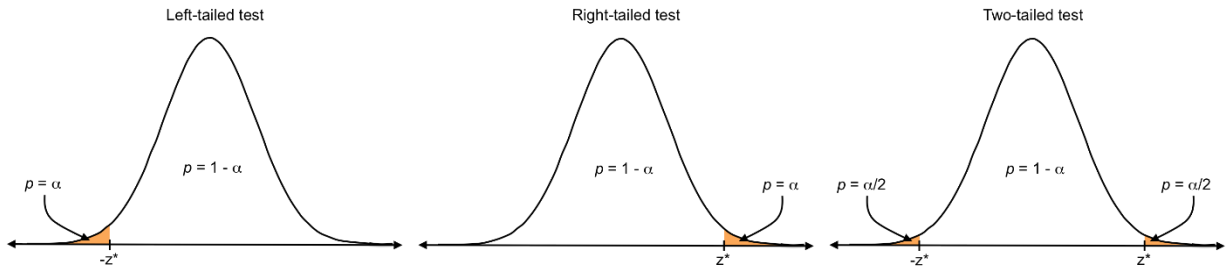


Figure 25: Z-critical values (z) mark the cutoff for the critical region, which adds to α.*

In Figure 25, the orange areas are the **critical regions**, and the cutoff is the **z-critical value**, which is based on the chosen α level. If a value or sample falls in the tail beyond the z-critical value, the results are considered significant.

We choose which test to run (left-tailed, right-tailed, or two-tailed) based on our hypothesis. If our hypothesis states that a particular value will be significantly less than the mean, we do a left-tailed test. If we're not sure, or if we merely speculate the value will be different, we do a two-tailed test.

Let's calculate the z-critical values for a two-tailed test at each α level. You can see from the z-table that for an α level of 0.05, a proportion of 0.025 is in each tail, and therefore the z-critical values are $\pm$1.96. We then say that if a value has a z-score less than -1.96 or greater than 1.96, it is statistically significant at $p < 0.05$.

What are the z-critical values for alpha levels of 0.01 and 0.001 for a two-tailed test? Well, 0.01 split between the two tails of the distribution indicates that 0.005 (0.5%) is in each tail. That means the cumulative probability up until the z-score marking the top 0.5% is 0.995. If you find that p = 0.995 in the body of the z-table, you will see that the corresponding z-score is about 2.58. So, $\pm$2.58 is the z-critical value for α = 0.01. Likewise, $\pm$3.27 is the z-critical value for α = 0.001. The most common alpha level used to test for significance is 0.05.

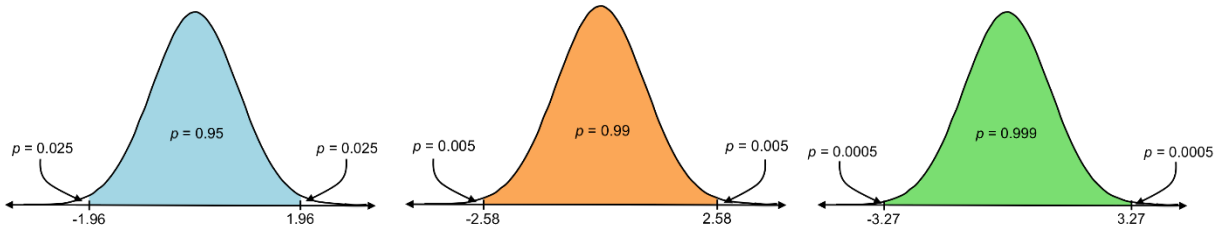


Figure 26: Z-critical values for alpha levels of 0.05, 0.01, and 0.001 are $\pm$1.96, $\pm$2.58, and $\pm$3.27.

Moving forward, we'll use this concept to estimate the population mean given a sample.

Chapter 5 One-Sample Z-Test

Calculate the likelihood of a random sample

Along with comparing individual values to others from the same normal distribution, we can compare a sample of values to other samples from the same population. This will help us determine if a particular sample we have collected is unlikely to occur. We test this essentially the same way we tested in Chapter 4. However, this time, because we have a sample, we look at where the mean of that sample falls in the distribution of means we would get from all other samples of the same size from that population. This distribution of sample means is called the **sampling distribution**.

If you collect a sample of size n from a population and calculate the mean ($\overline{x}_1$), then take another sample of size n from the same population and calculate the mean again ($\overline{x}_2$), and do this as many times as you possibly can so that each sample consists of a unique combination of values from the population, the sample means form a normal distribution. (Generally, sample sizes should be larger than 5.) Amazingly, it doesn't matter what the distribution of the population is. The population might have a bimodal, uniform, or skewed distribution, yet the sampling distribution will be normal (as long as the sample size—i.e. the number of values in each sample—is greater than 5). This phenomenon is called the **Central Limit Theorem**.

The sampling distribution is the distribution of all possible sample means of size n. Of course, this is theoretical; we can't possibly take every possible sample and find the mean. For example, if a population has 100,000 values and we have a sample size of 30, there would be

$$\binom{100{,}000}{30} = \frac{100{,}000!}{(100{,}000-30)!(30!)}$$

unique samples of size 30. This number would be ridiculously huge. However, by knowing what the sampling distribution would be if you could take every possible sample of size n, you can tell if something weird is happening with a particular sample.

The mean of the sampling distribution, which we'll call μ_M (i.e. the mean of the means), is equal to the population mean μ, and the standard deviation of the sampling distribution (σ_M) is equal to $\sigma / \sqrt{n}$ (the population standard deviation divided by the square root of the sample size). The standard deviation of the sampling distribution is called the **standard error (SE)**.

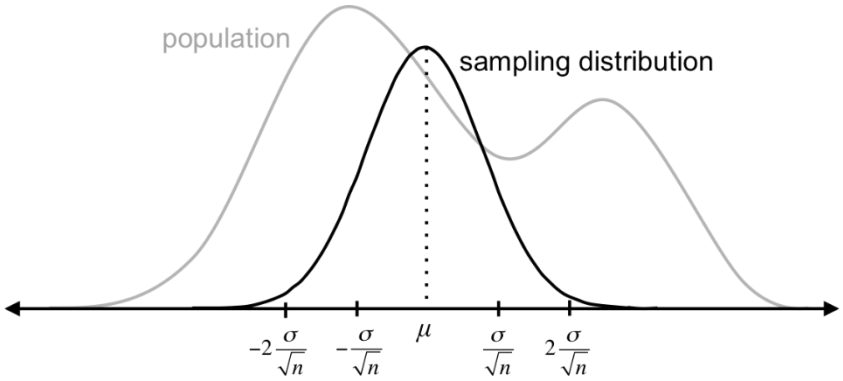


Figure 27: No matter the shape of the population, sampling distributions will follow a normal distribution as long as the sample size is greater than 5. The mean of the sampling distributions is equal to the mean of the population ($\mu = \mu_M$), and the standard deviation of the sampling distribution (the standard error) is the population standard deviation divided by the square root of the sample size ($\sigma_M = \sigma / \sqrt{n}$).

Since sampling distributions are normally distributed, about 68% of sample means fall within one standard error ($\sigma / \sqrt{n}$) of the population mean, and about 95% fall within two standard errors ($2\sigma / \sqrt{n}$). Therefore, it is very unlikely you'll get a random sample whose mean is more than two standard errors from the population mean in either direction. If this happens, something has probably been done to influence that sample.

When we conduct hypothesis testing for samples, we can test whether or not a particular sample mean is significantly different than the population mean μ, or from a particular value, or from another sample. The remainder of this chapter covers how to compare a sample mean to a specific value when we know population parameters μ and σ. This is called a **one-sample z-test**.

If the mean of a particular sample is significantly different from the mean of the population from which the sample was taken (μ), we assume that something has been done to influence the sample. If all values in the original population were similarly influenced, the entire population would shift to a new mean (μ_S), but the standard deviation would remain the same. (Some call the new population's mean μ_I for "influence" or "intervention.")

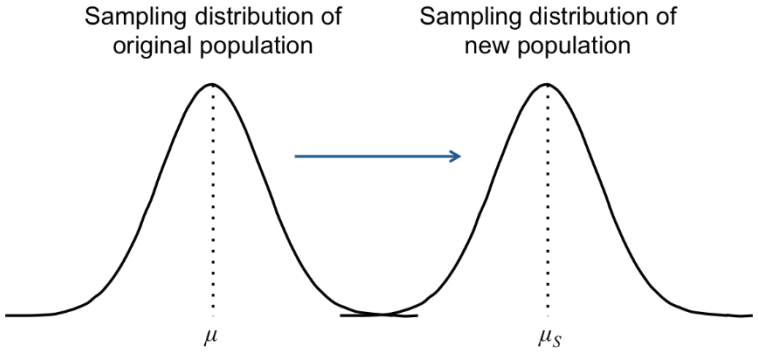


Figure 28: If the mean of a particular sample is significantly different from the mean of the population from which the sample was taken (μ), we assume that something has been done to influence the sample. If all values in the original population were similarly influenced, the entire population would shift to a new mean (μ_S), but the standard deviation would remain the same.

Our hypothesis test will produce one of two results: either the sample is not significantly different from μ, or it is. These two outcomes are called the null and alternative hypotheses. The **null hypothesis** states that the new population mean, based on the sample mean, is not significantly different from μ, and we notate this as such:

H_0: $\mu_S = \mu$

The **alternative hypothesis** states that the new population mean is significantly different, either by being significantly greater than the mean, significantly less than the mean, or one of the two:

H_a: $\mu_S < \mu$
H_a: $\mu_S > \mu$
H_a: $\mu_S \neq \mu$

When there is a significant difference, scientists will typically attempt to determine why that difference exists. They can do this through further quantitative analysis, qualitative research, or both.

We use the first two alternative hypotheses when we perform a one-tailed test (H_a: $\mu_S < \mu$ for a left-tailed test and H_a: $\mu_S > \mu$ for a right-tailed test) and the third alternative hypothesis for a two-tailed test.

Figures 29-31: The purple areas on each distribution depict the critical regions for $\alpha = 0.05$. If the sample mean falls in the critical region determined by the test, you're doing (one-tailed or two-tailed), the results are significant.

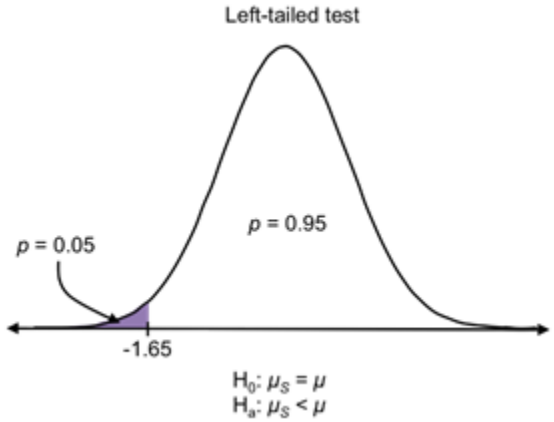


Figure 29

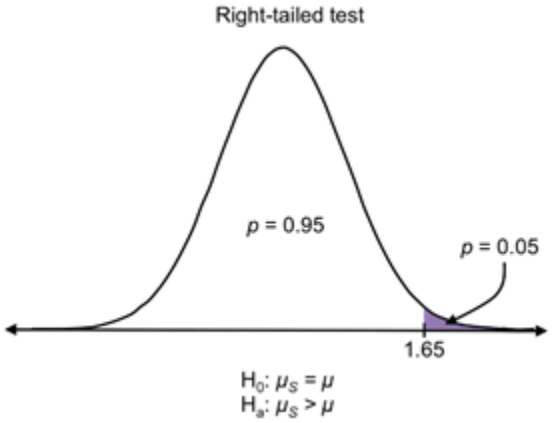


Figure 30

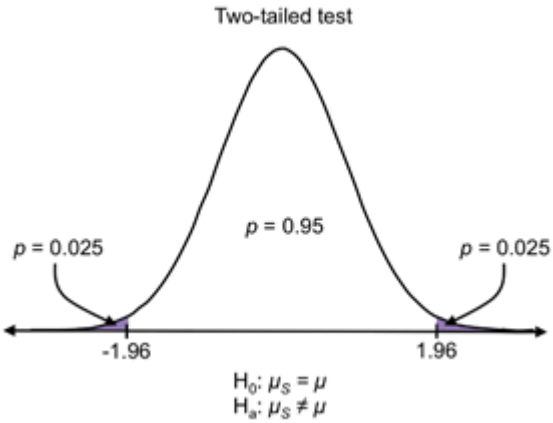


Figure 31

Example

Let's say a particular regional gym in the U.S., one with about 6,000 members, hires you to estimate the health of its membership so that it can boast of being an effective gym. You decide to use resting heart rate as an indicator of health, where the lower the resting heart rate, the healthier the individual.

You know from research that for all people in the United States aged 26-35, the mean resting heart rate is 73 with standard deviation 6. When you take a random sample of 50 gym members aged 26-35, you find that their average resting heart rate is 68. Can you say that members of this gym are healthier than the average person?

To answer this question, let's first write out the null and alternative hypotheses.

H_0: $\mu_S = 73$
H_a: $\mu_S \neq 73$

In order to determine whether to reject or accept the null hypothesis (that there is no significant difference in health between people at this gym and the general population), describe the sampling distribution of all samples of size 50 from the population of US adults aged 26-35. The mean is the same as the population mean (73) and the standard error is $\sigma / \sqrt{n} = 6 / \sqrt{50} = 0.85$. Where on this distribution does the sample mean (68) lie? In other words, how many standard errors is this particular sample mean from the population mean? We need to find the z-score of the sample mean:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{68 - 73}{0.85} = -5.88$$

You can see from the z-table that the probability of selecting a random sample from the population of 26-35-year-olds with a mean of 68 is about 0. Therefore, we can say that our results are statistically significant, with $p < 0.001$ (our smallest alpha level). We reject the null and conclude that the sample comes from a different population. In layman's terms, people at this gym are healthier than the average person.

(Of course, it could be that healthy people self-select to be members of this gym, and the gym isn't causing them to be healthier. To differentiate **correlation**—how two data sets change together—versus causation—whether or not one data set influences another—we could compare the members' resting heart rates before they joined the gym to their current resting heart rates. This involves another statistical test, which we'll examine in the next chapter.)

For now, let's focus on how to create a function that calculates the z-statistic in R given any mean and standard deviation. The following Code Listing creates this function, then chooses a random sample from the variable "income2011," then calculates the z-score for that sample.

Code Listing 5

```
z.test = function(a, mu, sigma){
       z.score = (mean(a) - mu) / (sigma.income / sqrt(length(a)))
       return(z.score)
    }  #creates a function (“z.test”) that will return the z-score
  “z.score” for specified values of “a,” “mu,” and “stdv”

> sample.income = sample(income2011, size=20, replace=FALSE, prob=NULL)
  #creates a sample of size 20 from income2011, without replacement, and
  calls it “sample.income”

> mu.income = mean(income2011) #specifies that “mu.income” is the mean of
  income2011

> sigma.income = sqrt(sum((income2011-
  mean(income2011))^2)/length(income2011))
  #treats income2011 as a population and specifies that “stdv.income” is
  equal to σ

> z.test(sample.income, mu.income, sigma.income) #calculates the z-
  statistic for sample.income

[1] 0.03153922
```

Note: In this example, the z-test returned 0.03 for the z-score. However, when you execute this, you will get a different result because the sample() function in R will generate a different random sample each time.

With the **`z.test()`** function you created, you can calculate the z-statistic for any three arguments you input. For example, if “a” is a set of specified values, **`z.test(a, 5, 3)`** will calculate the z-score for the mean of “a” on a sampling distribution that has mean 5 and standard deviation 3.

Find a range for the true mean

Once you determine that a sample is significantly unlikely to have occurred and therefore most likely comes from a new population with mean μ_S, you can determine a range in which you’re pretty sure μ_S lies. You can assume that this population has the same standard deviation as the original and that some kind of intervention has merely shifted the population one way or another.

Your best guess for μ_S is $\bar{x}$, but because you know that 95% of sample means are within 1.96 standard errors from the population mean, you can guess the true population mean μ_S will be within 1.96 standard errors from $\bar{x}$. This range is called a **95% confidence interval**, and as you saw before, 1.96 is the z-critical value that marks the middle 95% of the data.

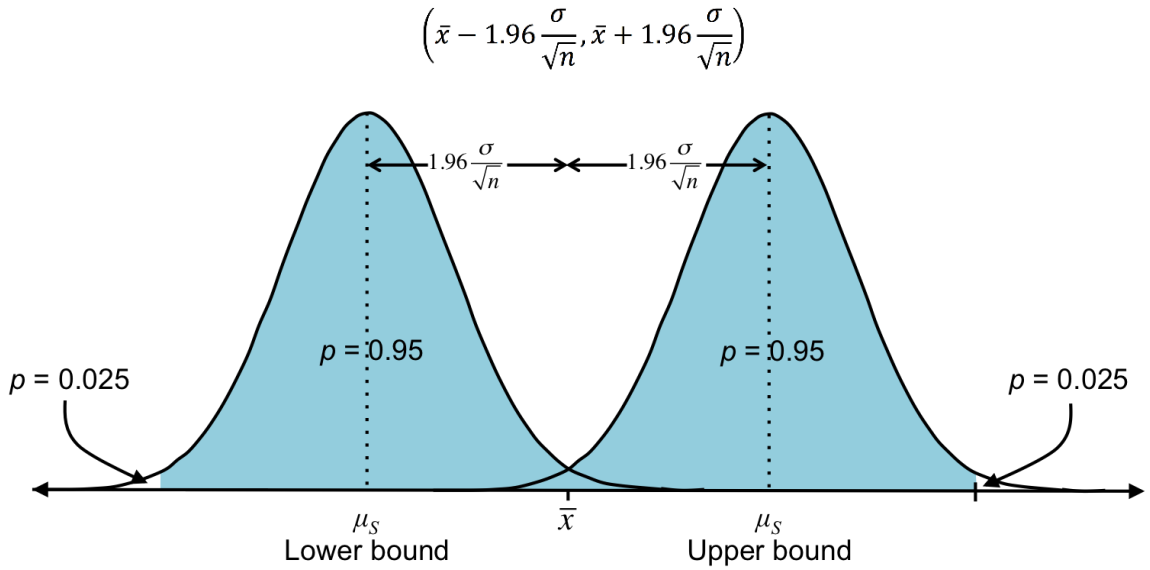


Figure 32: Each distribution above is the same sampling distribution consisting of all samples of size n. And $\bar{x}$ is the mean of a sample of size n. Most likely, this sample is one of the 95% within two standard errors of the population mean. Therefore, we can calculate a range in which we're pretty sure μ_S lies. This particular range is the 95% confidence interval.

Let's explore this in the context of our example.

Example

We want to estimate the mean resting heart rate of all members of this gym who are aged 26-35. This is the new population (rather than everyone in the US aged 26-35). The standard deviation of this new population remains the same as the old population: $\sigma = 6$.

We found that $\bar{x} = 68$. Most likely, this sample mean is one of the 95% of sample means that fall within 1.96 standard errors of the population mean. Assuming it is, μ_S would fall between $68 - 1.96\frac{6}{\sqrt{50}} = 66.34$ and $68 + 1.96\frac{6}{\sqrt{50}} = 69.66$.

Other confidence intervals

We could also come up with a broader range, such as a 98% confidence interval. In this case, we would be even more certain that μ_S is in this range. You can find this interval by first determining the z-critical values that mark the middle 98% of the data. The z-table tells us these values are ±2.33 (since 1% is in each tail). So, you're pretty sure that the sample lies within 2.33 standard errors of the population mean μ_S, which tells us that the 98% confidence interval for μ_S: $\left(\bar{x} - 2.33\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.33\frac{\sigma}{\sqrt{n}}\right)$

In general terms, the bounds of a C% confidence interval are

$$\left(\bar{x} - z^*\left(\frac{\sigma}{\sqrt{n}}\right), \bar{x} + z^*\left(\frac{\sigma}{\sqrt{n}}\right)\right)$$

where z* is the z-critical value marking the lower and upper bounds of the middle C% of the distribution.

The following R code creates functions to calculate the lower and upper bounds of a 95% confidence interval. The inputs “sample.income” and “sigma.income” continue from Code Listing 5 and are used to calculate a 95% confidence interval based on the random sample “sample.income.”

Code Listing 6

```
> ci.lower = function(a, sigma){
        lower.bound = mean(a) - 1.96*(sigma/sqrt(length(a)))
        return(lower.bound)
        } #creates a function (“ci.lower”) that will return the lower
  bound (“lower.bound”) of a 95% confidence interval for specified values
  of “a” and “sigma”

> ci.upper = function(a, sigma){
        upper.bound = mean(a) + 1.96*(sigma/sqrt(length(a)))
        return(upper.bound)
        } #creates a function (“ci.upper”) that will return the upper
  bound (“upper.bound”) of a 95% confidence interval for specified values
  of “a” and “sigma”

> ci.lower(sample.income, sigma.income) #returns the lower bound

[1] 16723.69

> ci.upper(sample.income, sigma.income) #returns the upper bound

[1] 38226.31
```

The sample “sample.income” indicates that the true mean of the population it comes from is between 16,723.69 and 38,226.31. Note that the mean of “income2011” is 27,302. The fact that 27,302 lands smack in the middle of this confidence interval makes sense given that “sample.income” has a z-score of 0.03, which is found in the middle of the sampling distribution.

Calculate the likelihood of a proportion

Along with determining whether or not a sample mean is likely to occur, sometimes we also want to determine if a proportion is likely to occur. For example, if a certain diet plan says that half of people who begin this plan start losing weight after one week and we find that of 30 people, only 11 lost weight after one week, we can execute a test to determine if the proportion who lost weight is significantly less than 50% (which is what we would expect).

We denote our expected proportion as $\hat{p}$ and the observed outcome as p. So, our general null and alternative hypotheses are:

H_0: $p = \hat{p}$
H_a: $p \neq \hat{p}$

In this example, our null and alternative hypotheses are:

H_0: p = 0.5
H_a: p ≠ 0.5

To test for significance, we'll again think about our expected, theoretical sampling distribution. If we take every possible sample of size n from our expected population (in this example, those who participate in this diet plan) and find the proportion of each sample that obtained the result we're interested in (in this case, those who lost weight after a week), the distribution of sample proportions would be expected to have mean $\hat{p}$ and this standard deviation:

$$\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

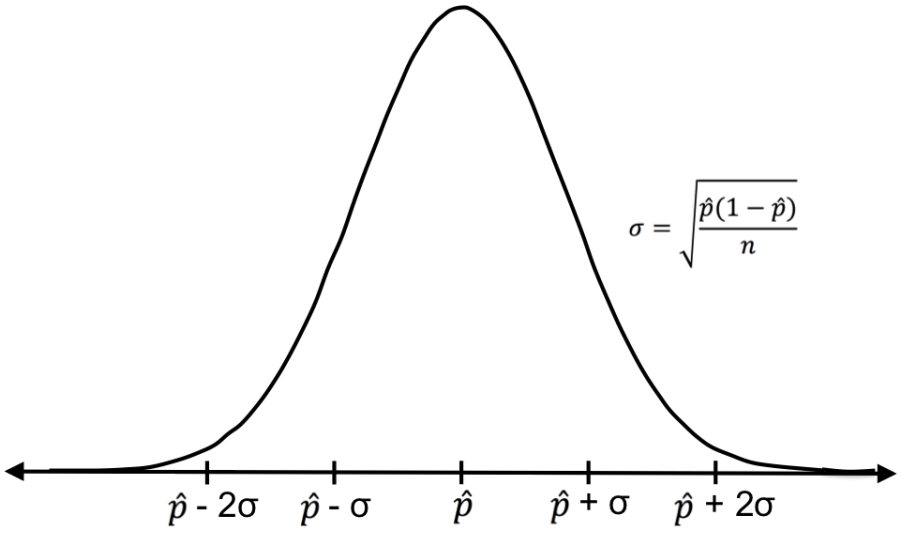


Figure 33: The sampling distribution for sample proportions has mean $\hat{p}$ (which is equal to the population proportion) and standard deviation $\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Now that we know the mean and standard deviation of the sampling distribution, we can determine where the observed outcome (p = 11/30 = 0.367) falls on this distribution. This is another z-test.

$$z = \frac{p - \hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

In our example,

$$z = \frac{0.367 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{30}}} = -1.46$$

the result is greater than the z-critical value on the left tail of -1.96. Therefore, the results are not significant, i.e. the proportion of those who lost weight is not significantly less than 0.5. This diet plan's claim could still be true!

Let's now preform a test for proportions using the NCES data. Perhaps we want to know if the proportion of students who worked while in school is significantly more or less than 0.5. Again, we'll create a function that returns the z-statistic.

Code Listing 7

```
> p.test = function(a, phat){
  z.score = (mean(a)-phat)/sqrt(phat*(1-phat)/length(a))
  return(z.score)
  } #creates a function ("p.test") that will return the z-score "z.score"
  for specified values of "a" and "phat"

> p.test(work, 0.5) #calculates the z-statistic for the proportion of
  students who worked compared to the proportion 0.5

[1] -20.93313
```

Now you can use the **`p.test()`** function for any set of binary data to test whether the proportion of one of the values significantly differs from any specified proportion—in this case 0.5. As you can see from the z-score of -20.9, the result of this particular test reveals that the proportion of students who worked during high school is significantly less than 0.5.

By now, you should be comfortable calculating probabilities under the normal curve and using these probabilities to determine if a value, sample, or proportion is out of the ordinary. In the next chapter, we'll look at which kinds of test to execute when you don't know original population parameters μ and σ.

Chapter 6 T-Tests

Hypothesis test when population parameters are unknown

In this chapter, you'll determine if a sample mean is significantly different from a particular value (which we'll call μ_0) or another sample when we don't know population parameters. To do this, we'll use the same concept as in Chapter 5: determine how many standard errors are separating $\bar{x}$ from μ_0 or $\bar{x}$ and the mean of the other sample. The procedure is exactly the same—only the calculation of standard error changes.

If we have a sample but don't know population parameters, we have to make conjectures about the population based on s (sample standard deviation) and $\bar{x}$ (mean).

$$\sigma \approx s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

We use the sample standard deviation, s, to approximate the population standard deviation, σ. (Remember, we divide the sum of squares by n-1 rather than n in order to make the result slightly bigger and therefore to better approximate σ.)

Consequently, we're also approximating the standard error: s / √n. Because we use s to approximate the standard error, we will have more error and therefore we will use a different kind of distribution that has thicker tails: the **t-distribution**.

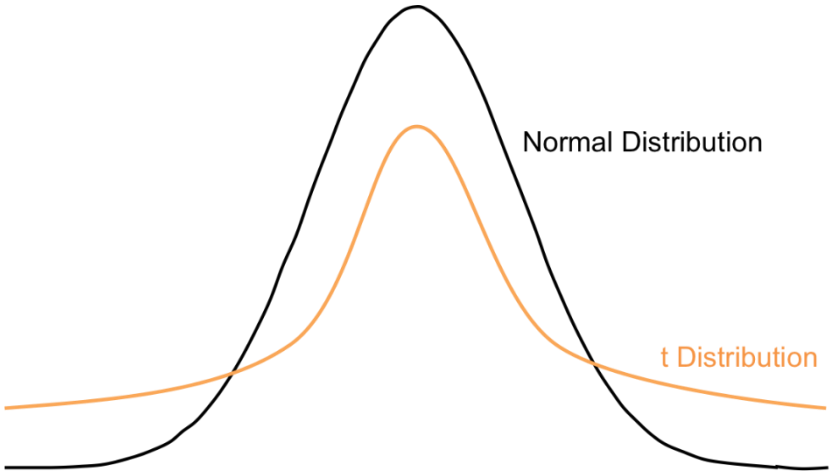


Figure 34: The t-distribution has fatter tails to compensate for the greater error involved in calculating the standard error, which uses s to approximate σ.

Because we're using an entirely different distribution that accounts for greater error, these statistical tests are called **t-tests**.

Again, we integrate the PDF of the t-distribution to calculate probabilities. And also again, we have a table (this time the **t-table**) that provides the probabilities for us. A few things differ between this table and the z-table:

- The t-table asks for the **degrees of freedom**, which is equal to n – 1.
- The body of the t-table gives **t-critical values** rather than probabilities (which are the column headers).
- The t-table provides t-critical values for both one-tailed and two-tailed tests.

Tip: More detail on degrees of freedom can be found in Street-Smart Stats, Chapter 10.

The t-table supplies t-critical values because we no longer care about probabilities under this curve the way we do with standard distributions. We only care about whether or not our sample mean falls within the critical region, and to learn this we simply need to compare the distance in terms of standard errors (i.e. the **t-statistic**) with the t-critical value marking our chosen alpha level.

We denote the t-critical value $t_{(\alpha, df)}$ to specify the alpha level we're using and the degrees of freedom. Here is the t-statistic:

$$t = \frac{\bar{x} - \mu_0}{SE}$$

If we perform a two-tailed test (*high or low tails*) at $\alpha = 0.05$, and our sample size is n = 37, then $t_{(0.05, 36)} = 2.021$ (with degrees of freedom rounded from 36 to 40). Therefore, if the t-statistic is either less than -2.021 or greater than 2.021, t falls in the critical region and our results are significant.

We'll cover two types of t-tests in this chapter:

1. One-sample t-tests in which we compare a sample to a constant, denoted μ_0.
2. Two-sample t-tests in which we compare two samples with each other.

One-sample t-tests

We perform a one-sample t-test when we want to know if our sample and a particular value (μ_0) are likely to belong to the same population. In other words, if μ_0 were a sample mean, would it likely be in the same sampling distribution as $\bar{x}$?

A t-test answers this question by estimating the standard error and then determining the number of standard errors that separate $\bar{x}$ and μ_0. We use the t-table to determine if the probability of those errors being this distance apart is less than our alpha level.

In this case, our t-statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

with null and alternative hypotheses:

Left-tailed test	Right-tailed test	Two-tailed test
H_0: $\mu = \mu_0$ H_a: $\mu < \mu_0$	H_0: $\mu = \mu_0$ H_a: $\mu > \mu_0$	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$

Note that μ_0 is a constant.

Example

A major technology company has just released a new product. The business development team has decided that on a scale of 1 to 10 (10 being highest satisfaction), their goal is for customer satisfaction to achieve more than 8. The R&D department decides to perform a one-sample t-test to determine if customer satisfaction so far is significantly more than 8. They send out a survey to a random sample of 50 customers who bought the product, and they find the average satisfaction score is 8.7 with sample standard deviation 1.6. Does it appear that most people will rate their satisfaction above 8?

In this case, we'll do a one-sample t-test because we want to determine if the customer satisfaction score is greater than 8 rather than simply different from 8. Therefore, our null and alternative hypotheses are:

H_0: $\mu_S = 8$
H_a: $\mu_S > 8$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{8.7 - 8}{1.6/\sqrt{50}} = 3.09$$

Now, we must compare this t-statistic with the t-critical value, $t_{(0.05, 49)}$. The t-table tells us that for a one-tailed test, $t_{(0.05, 40)} = 1.684$. (Note that $t_{(0.05, 60)} = 1.671$, so for our particular sample with df = 49, $t_{(0.05, 49)}$ will fall between 1.671 and 1.684.) Because the t-statistic is greater than the t-critical value, our results are significant, and we can say that customer satisfaction is statistically significantly greater than 8 and we reject the null. (Recall that the null hypothesis states that the results are not significant, so rejecting the null means we've concluded that there is a significant difference between the observed customer satisfaction and the goal of 8.)

Let's do a one-sample t-test in R for the variable "income2011" of our NCES data.

Code Listing 8

```
> t.test(income2011, mu = 40000) #determine if the mean income is
significantly different from $40,000

      One Sample t-test

data:  income2011
t = -47.0042, df = 8246, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 40000
95 percent confidence interval:
 26772.44 27831.55
sample estimates:
mean of x
    27302
```

This t-test takes **`mean(income2011)-40,000`**, so if the t-statistic is negative, **`mean(income2011)`** is less than 40,000. You should find that t = -47, indicating that the true population mean is a lot less.

You've just learned how to compare a sample to a particular value, μ_0. We also use t-tests to determine if two samples are significantly different and therefore most likely come from two different populations. In this case, we would do a two-sample t-test.

Two-sample t-tests

There are two types of two-sample t-tests: dependent-samples t-tests and independent-samples t-tests. We use dependent samples when measurements are taken from the same subjects. For example, we might use a dependent-samples t-test to determine if there is a significant difference between:

- Children's heights at age 8 and age 10.
- Students' scores on a pre-assessment and post-assessment for a course.
- The effectiveness of two different sleeping pills on the same people.

This method controls for individual differences, in effect holding them constant to determine if any difference is most likely due to the intervention (in the examples above, the interventions are age, course materials, and the sleeping pill).

We use independent samples when subjects differ between the two groups. For example, we might use this test to determine if a significant difference exists between:

- Different countries' carbon emissions.
- Men's and women's wages.
- Flight costs in January vs. August.

Independent samples no longer have the benefit of holding individual subjects constant. In order to determine if there is a significant difference between groups, samples need to be random.

To help us think about each two-sample test, let's assign symbols representing each sample's statistics.

Sample 1
Mean = $\overline{x}_1$
Sample standard deviation = s_1

Sample 2
Mean = $\overline{x}_2$
Sample standard deviation = s_2

Our null and alternative hypotheses are:

Left-tailed test	Right-tailed test	Two-tailed test
H_0: $\mu_1 = \mu_2$ H_a: $\mu_1 < \mu_2$	H_0: $\mu_1 = \mu_2$ H_a: $\mu_1 > \mu_2$	H_0: $\mu_1 = \mu_2$ H_a: $\mu_1 \neq \mu_2$

We can rewrite them as:

Left-tailed test	Right-tailed test	Two-tailed test
H_0: $\mu_1 - \mu_2 = 0$ H_a: $\mu_1 - \mu_2 < 0$	H_0: $\mu_1 - \mu_2 = 0$ H_a: $\mu_1 - \mu_2 > 0$	H_0: $\mu_1 - \mu_2 = 0$ H_a: $\mu_1 - \mu_2 \neq 0$

Here the difference between the two populations is significantly less than, greater than, or different from 0.

Dependent samples t-test

The dependent-samples t-test is almost exactly like the one-sample t-test. The only thing that changes is that we take the difference between each value measured from each subject, and that group of differences becomes our sample. We then test to see if this difference is significantly different than 0.

Example

Let's say we want to determine whether an online course was effective in improving students' abilities. To measure this, students were given a standard test before they went through the course, and they took a similar test after completing the course

Our data looks something like this:

Student	Pre-test score	Post-test score	Difference
1	70	73	3
2	64	65	1
3	69	63	-6
…	…	…	…
34	82	88	6

The differences (which we'll denote as d_i) become our sample. Now we'll proceed as we would for a one-sample t-test, with $\mu_0 = 0$.

Let's say that upon taking each difference d_i, we find the following mean and sample standard deviation:

$\bar{d}_i = 4.2$
$s_i = 3.1$

We can now find our t-statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.2 - 0}{3.1/\sqrt{34}} = 7.9$$

This is much greater than the t-critical value $t_{(0.05,\ 33)} \approx 2.042$ for a two-tailed test. Therefore, we reject the null and conclude that the online course improved scores.

Now we'll perform a dependent-samples t-test in R. This time we won't use the NCES data because there isn't necessarily anything we would want to perform a t-test on, so we'll use a different data set. Let's say we want to know whether stocks significantly dropped following the Greek bank closures on June 29, 2015, so we analyze the previous closing price (on June 28) and the closing price on June 29 for 15 companies.

First, visit turnthewheel.org/street-smart-stats/afterward and click "Stock data set" (the second link under "Resources") to view the data in a Google spreadsheet. Download the data as a .csv, rename it "stocks.csv," and save the file to your working directory. The code follows.

Code Listing 9

```
> stocks = read.csv(file = "stocks.csv", head = TRUE, sep = ",") #input
the data into R
```

```
> attach(stocks) #allow R to recognize variable names

> t.test(today, yesterday, paired=TRUE) #tests whether the mean
difference (today – yesterday) is significantly different than 0

      Paired t-test

data:  today and yesterday
t = -6.4302, df = 14, p-value = 1.573e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6215956 -0.8104044
sample estimates:
mean of the differences
                 -1.216
```

The **`t.test()`** function returns a t-statistic for the first argument minus the second argument. In other words, if the t-statistic is negative, that means the first argument is less than the second. To ensure you're doing a dependent-samples t-test, type **`paired=TRUE`**.

In this example, we see that t = -6.4302, which is statistically significant—i.e. the average stock price dropped more than $1 between yesterday and today.

> ***Note: This example has been simplified for the purpose of explaining how to apply a t-test. However, a robust t-test must use a larger sample size or normally distributed data. Otherwise, comparing two data sets doesn't make much sense because there will be too much error and uncertainty in the way they are distributed (i.e. a $1 drop in price may mean something completely different for one stock versus another).***

Independent-samples t-test

Things get a bit more complicated with independent-sample t-tests because we can't simply subtract the values as we can with a dependent-sample test. Not only do we have different sample sizes, but we also have to account for the standard deviations of both samples rather than simply taking the standard deviation of the differences.

We can think about whether or not two samples are significantly different (i.e. they could come from the same population) in the context of confidence intervals. We can use each sample to determine a range in which we're pretty sure each population mean lies:

$$\bar{x} \pm t_{(\alpha,\ df)}\left(\frac{s}{\sqrt{n}}\right)$$

If these intervals overlap a lot, the two samples might have been taken from the same population, and it's due to error that the samples are different (because every sample taken from a population will most likely include different values than the next). However, we also have to consider the standard deviation of each population distribution. The greater the standard deviation of each distribution, the more likely the distributions will overlap. And the more they overlap, the more likely the two samples came from the same population.

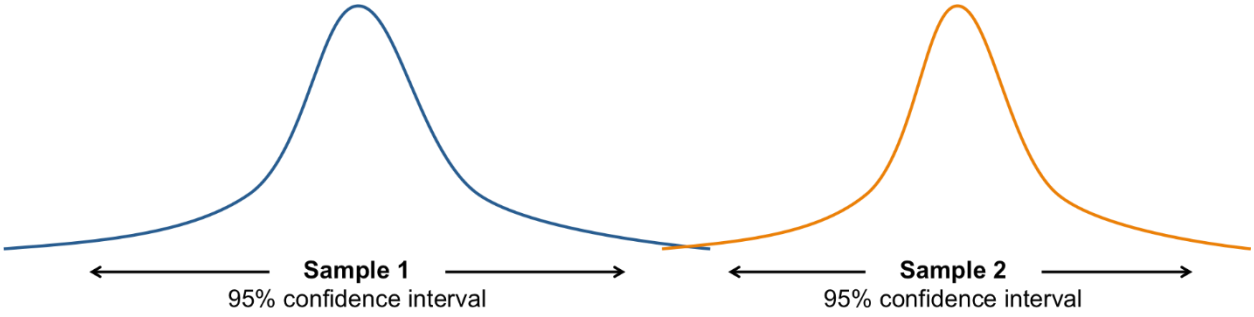


Figure 35: Most likely, these samples come from two different populations.

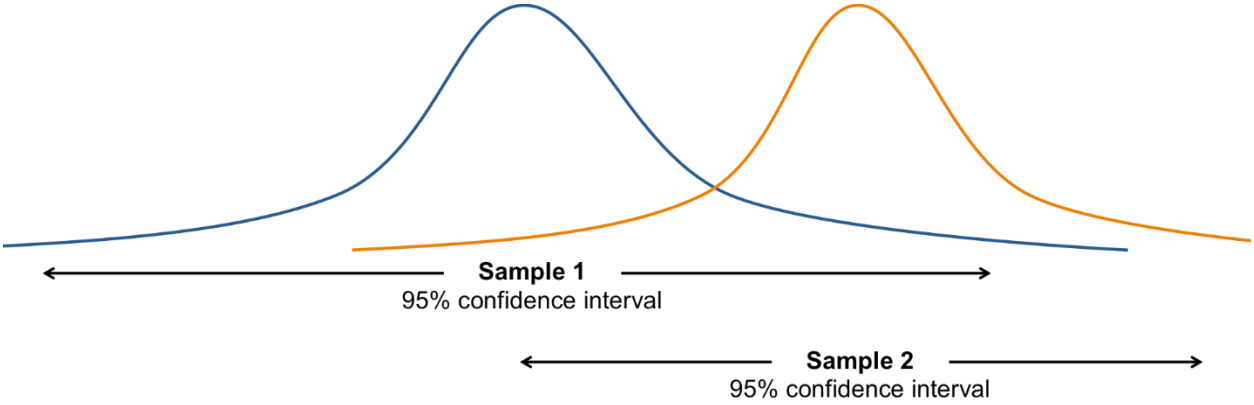


Figure 36: These samples might come from the same population, and the sample means differ simply due to chance.

We use a t-test to decide if the samples are statistically different or if they differ due to chance. We have two independent samples, each with their own standard deviations, which means we need to pool the standard deviations together to calculate the standard error. The t-statistic becomes:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Essentially, this determines whether or not the difference between sample means is significantly different than 0 (similar to the one-sample t-test):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Because we have two different sample sizes, the degrees of freedom is the sum of the individual degrees of freedom: $n_1 + n_2 - 2$.

Example

Let's say we want to test whether or not a gender wage gap exists for independent contractors. If we perform a two-tailed test, our null and alternative hypotheses are:

H_0: $\mu_M - \mu_F = 0$
H_a: $\mu_M - \mu_F \neq 0$

Here μ_M is the population of male contractors' hourly rates, and μ_F is the population of female contractors' hourly rates.

We take a random sample of 17 male and 15 female independent contractors and find each person's hourly rate. Then we calculate each mean and sample standard deviation.

Males
$\bar{x}_M$ = \$37
s_M = \$18

Females
$\bar{x}_F$ = \$33
s_F = \$12

Therefore, the t-statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} = \frac{37 - 33}{\sqrt{\frac{18^2}{17} + \frac{12^2}{15}}} = 0.747$$

with df = 30.

The t-table tells us that $t_{(0.05, 30)} = 2.042$ for a two-tailed test. In other words, the t-critical value marking the bottom 2.5% and top 2.5% is ± 2.042. In this case, we fail to reject the null, and we conclude that there is no significant difference between male and female independent-contractor hourly rates for these two populations.

> ***Tip: We never say that we "accept" the null because we don't truly know if the null hypothesis is true. To know this, we would need to know information for the entire population. Instead, we're basing conclusions off of a sample that only supports the conclusion that we don't have enough information to reject the null just yet. So, the proper way to write our conclusion is that we "fail to reject the null."***

There are ways to do an independent-samples t-test in R:

- **`t.test(a, b)`**
  In this case, "a" and "b" are arguments for a set of numerical values. For example, a =

{1, 3, 4, 5, 5, 3, 7} and b = {4, 3, 4, 2, 1}. This t-test tells us if the means of "a" and "b" are significantly different.

- **t.test(a ~ b)**
  This code tells us how values of "a" differ by values of "b," where "b" is categorical.

We can use the code t.test(a ~ b) for our NCES data to determine if people who worked during high school had a higher income in 2011 than those who did not work. Let's explore this data with a t-test on 2011 income based on whether or not they worked in high school. Since socioeconomic status (SES) may have played a role in people's decision to work during high school, let's further explore the data with another t-test on SES based on whether or not they worked.

Our null and alternative hypotheses for a two-tailed test are:

H_0: $\mu_{income:\ work} = \mu_{income:\ did\ not\ work}$
H_a: $\mu_{income:\ work} \neq \mu_{income:\ did\ not\ work}$

H_0: $\mu_{SES:\ work} = \mu_{SES:\ did\ not\ work}$
H_a: $\mu_{SES:\ work} \neq \mu_{SES:\ did\ not\ work}$

Code Listing 10

```
> t.test(income2011 ~ work) #test for a significant difference in 2011
income based on whether students worked during high school

      Welch Two Sample t-test

data:  income2011 by work
t = -3.5501, df = 6749.586, p-value = 0.0003877
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3055.1956  -881.4297
sample estimates:
mean in group 0 mean in group 1
       26544.70        28513.01

> t.test(ses ~ work) #tests whether the mean difference (today –
yesterday) is significantly different than 0

      Welch Two Sample t-test

data:  ses by work
t = -0.1809, df = 6996.377, p-value = 0.8565
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03510958  0.02917851
sample estimates:
mean in group 0 mean in group 1
```

```
      0.1230666       0.1260321
```

Our results show that students who worked during high school (coded “1”) had higher incomes in 2011 and that the results are significant at $p < 0.001$. Therefore, in this case we reject the null hypothesis and can conclude that students who worked in high school later earned higher incomes.

However, our second t-test reveals that SES was not significantly different between those who worked and didn’t work during high school, so SES was not a factor in whether or not students chose to work. These findings support the first hypothesis: that students who worked during high school had a stronger work ethic, and as a result they made more money 10 years later.

Now that you’ve learned how to tell if two samples are significantly different, you’ll learn how to discern if any two samples in a group of three or more samples are significantly different.

Chapter 7 ANOVA

Test for differences between three or more samples

When we have three or more samples and we want to determine if any two of them are significantly different from each other, we utilize a statistical test called **analysis of variance (ANOVA)**. This test analyzes how a **dependent variable** differs based on one or more **independent variables**. For example, we may want to know if scores on a standardized math test (the dependent variable) are significantly different between students who go to School A, School B, and School C (where school is the independent variable).

The dependent variable should be continuous (math test scores are continuous), be approximately normally distributed, and have homogenous variances for the values for each of the groups. The independent variable should be categorical with mutually exclusive values (where "School A," "School B," and "School C" are the different values). The independent variable should also be approximately normally distributed.

This chapter covers both one-way and two-way ANOVA. In one-way ANOVA, there is one independent variable; in two-way ANOVA there are two independent variables.

Compare samples based on one factor

The null and alternative hypotheses are:

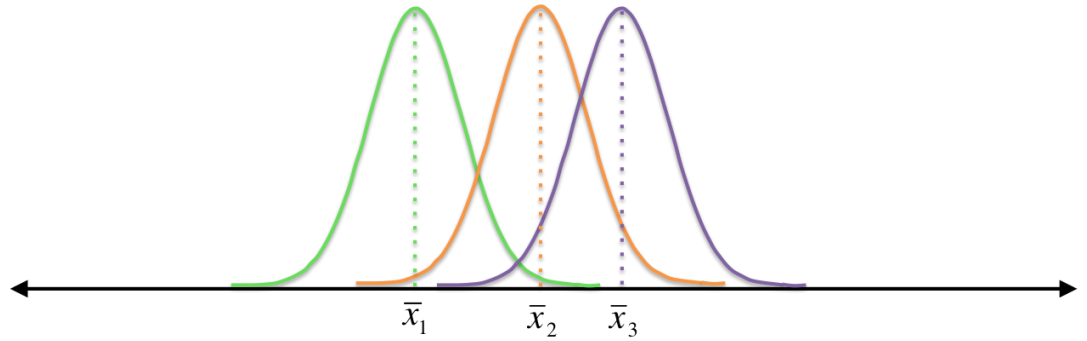
H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H_a: at least two population means are significantly different

where there are k samples.

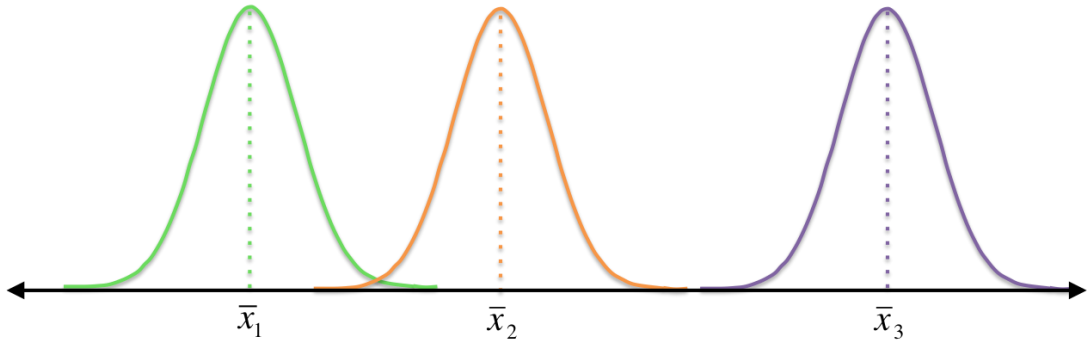
This test involves measuring **between-group variability**—essentially, the variance of the sample means—and dividing between-group variability by a measure of **within-group variability**—a combined measure of each sample's variance. The resulting quotient is the F-statistic.

F = (between-group variability) / (within-group variability)

The greater F is, the more likely at least two populations are significantly different. If you think about this quotient, greater between-group variability indicates that the sample means are spread out farther from each other, which implies they're more likely to be significantly different.



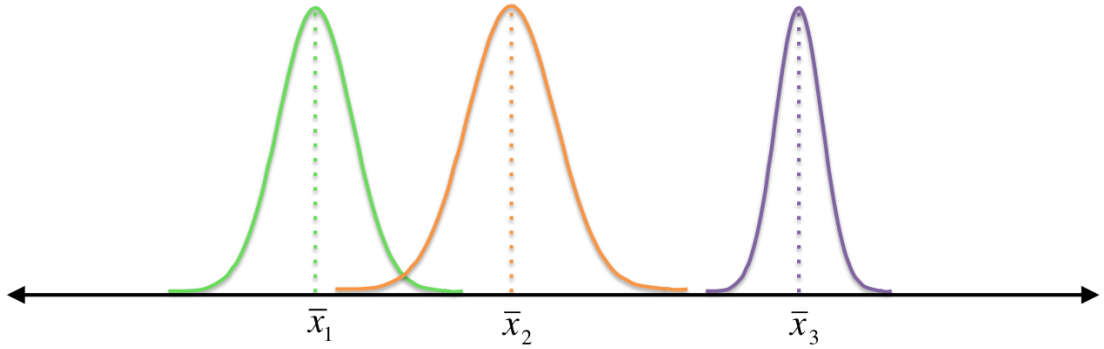
Smaller between-group variability



Greater between-group variability

Figure 37: In the previous figure, within-group variability is the same, but between-group variability changes. As between-group variability increases, samples are more likely to be significantly different and the F-statistic increases.

On the other hand, greater within-group variability indicates that the standard deviation of each sample is greater, which implies the samples are less likely to be significantly different.



Smaller within-group variability

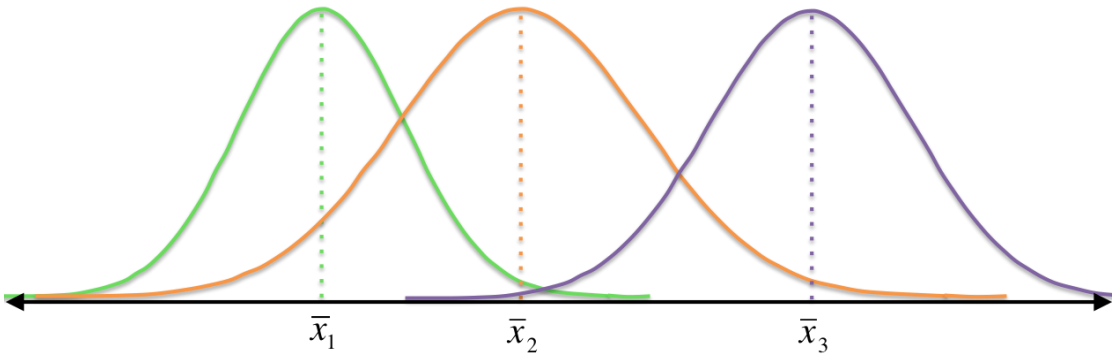


Figure 38: In this figure, between-group variability is the same, but within-group variability changes. As within-group variability increases, samples are less likely to be significantly different and the F-statistic decreases.

So, the greater the numerator (between-group variability) and the smaller the denominator (within-group variability), the greater the F statistic will be. Let's calculate each of these in turn.

Between-group variability

Measuring the spread of sample means is just like calculating the variance. First, we find the **grand mean** ($\bar{x}_G$), which is the sum of all values from each sample divided by the sum of each sample size.

$$\bar{x}_G = \frac{\Sigma x_{1i} + \Sigma x_{2i} + \cdots + \Sigma x_{ki}}{n_1 + n_2 + \cdots + n_k} = \frac{\Sigma\Sigma x_{ji}}{N}$$

In the first expression, x_{1i} represents all values from sample 1, x_{2i} represents all values from sample 2, etc., while n_1 is the size of sample 1, n_2 is the size of sample 2, etc., for k samples; in the second expression, j is the sample number that ranges from 1 to k, while i is the value number in each respective sample.

To calculate between-group variability:

1. Calculate the deviation of each sample mean from the grand mean: $(\bar{x}_j - \bar{x}_G)$.
2. Square each deviation: $(\bar{x}_j - \bar{x}_G)^2$.
3. Multiply each squared deviation by the sample size of that respective sample to weigh the squared deviation by the sample size: $n_j(\bar{x}_j - \bar{x}_G)^2$.
4. Sum each weighted squared deviation: $\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x}_G)^2$.
   This gives us the **sum-of-squares for between-group variability ($SS_{between}$)**.

5. Find the average squared deviation by dividing $SS_{between}$ by the degrees of freedom, where $df_{between} = k - 1$. This quotient is known as the mean square for between-group variability ($MS_{between}$), and this is our final measure of between-group variability.

$$\text{Between-group variability} = \frac{SS_{between}}{df_{between}} = MS_{between} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x}_G)^2}{k-1}$$

Within-group variability

This is a measure of error that combines the error within each individual sample. To calculate within-group variability:

1. Calculate the deviation of each value from the mean of the sample it comes from: $(x_{ji} - \bar{x}_j)$.

2. Square each deviation: $(x_{ji} - \bar{x}_j)^2$.

3. Sum the squared deviations: $\Sigma\Sigma(x_{ji} - \bar{x}_j)^2$.
   We now have the **sum of squares for within-group variability (SS_{within})**.

4. Find the average squared deviation of each value from its respective sample mean by dividing SS_{within} by df_{within} (the sum of the degrees of freedom of each sample), where $df_{within} = (n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1) = n_1 + n_2 + \ldots + n_k - k = N - k$. This quotient is the mean square for within-group variability (MS_{within}) and is our final measure of within-group variability.

$$\text{Within-group variability} = \frac{SS_{within}}{df_{within}} = MS_{within} = \frac{\Sigma\Sigma(x_{ji} - \bar{x}_j)^2}{N-k}$$

Total variability

If we take the deviation of each individual value from the grand mean, square each deviation, and find the sum of squared deviations, we get the total sum-of-squares (SS_{total}), a measure of the total variability. What's cool is that SS_{total} is the sum of $SS_{between}$ and SS_{within}.

$SS_{between} + SS_{within} = SS_{total}$

We often organize all our calculations in an ANOVA table.

Table 5: ANOVA tables help organize variability calculations.

	SS	**df**	**MS**	**F**
Factor	$SS_{between}$	$df_{between}$	$MS_{between} = \frac{SS_{between}}{df_{between}}$	$F = \frac{MS_{between}}{MS_{within}}$
Error	SS_{within}	df_{within}	$MS_{within} = \frac{SS_{within}}{df_{within}}$	
Total	$SS_{between} + SS_{within}$	$df_{between} + df_{within}$		

The F-distribution

As with z- and t-statistics, F-statistics follow a specific distribution. In our calculation of between- and within-group variability, we (essentially) found average squared deviations. Therefore, F can never be negative, and the distribution lies along the positive x-axis.

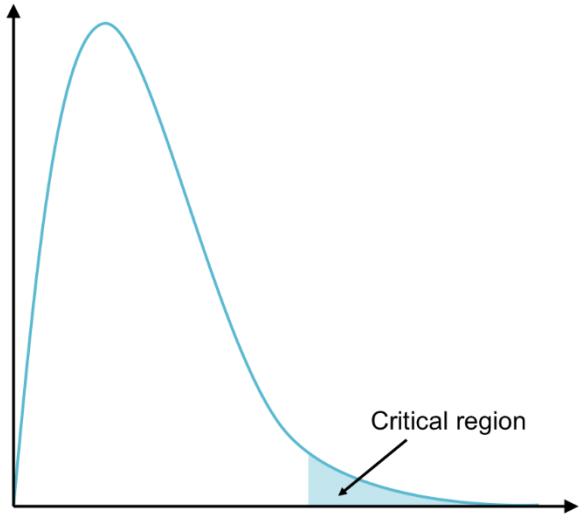


Figure 39: The F distribution has one tail on the right (on the positive x-axis). The critical region lies in this tail.

This also makes sense in the context of what we're testing for—we simply want to know if at least two populations are significantly different, not if one is significantly less than or greater than another.

Once we've calculated the F-statistic, we use the F-table to determine the **F-critical value, $F(\alpha, df_{between}, df_{within})$**. You'll see in the F-table that the column headers are $df_{between}$ and the row headers are df_{within}. A specific F-table exists for each alpha level.

Example

As a digital marketer, you always try to determine the most effective means of advertising online. Let's say you've placed ads with a number of online publications, in various locations on each web page, and you decide to perform ANOVA to determine if different locations have better click-through rates (CTR, which is the ratio between the percentage of users who click the ads and the total site visitors). The three most common ad locations are on the top, middle, and sides of the page. In your research, you find the CTR of more than 600 ads.

Table 6: This ANOVA table helps organize between-group and within-group variability calculations.

Top	Middle	Sides
n_T = 235 $\bar{x}_T$ = 35% s_T = 3.2%	n_M = 169 $\bar{x}_M$ = 28% s_M = 4.3%	n_S = 210 $\bar{x}_S$ = 38% s_S = 2.5%

This gives us everything we need to calculate between-group and within-group variability.

Between-group variability:

$$\bar{x}_G = \frac{\Sigma x_{Ti} + \Sigma x_{Mi} + \Sigma x_{Si}}{n_T + n_M + n_S}$$

We know that:

$$\bar{x}_T = \frac{\Sigma x_{Ti}}{n_T} \overset{therefore}{\Longrightarrow} \Sigma x_{Ti} = \bar{x}_T n_T$$

$$\bar{x}_M = \frac{\Sigma x_{Mi}}{n_M} \overset{therefore}{\Longrightarrow} \Sigma x_{Mi} = \bar{x}_M n_M$$

$$\bar{x}_S = \frac{\Sigma x_{Si}}{n_S} \overset{therefore}{\Longrightarrow} \Sigma x_{Si} = \bar{x}_S n_S$$

And therefore,

$$\bar{x}_G = \frac{\bar{x}_T n_T + \bar{x}_M n_M + \bar{x}_S n_S}{n_T + n_M + n_S} = \frac{(35\%)(235) + (28\%)(169) + (38\%)(210)}{235 + 169 + 210} = 34.1\%$$

This leads us to $SS_{between} = (35 - 34.1)^2 + (28 - 34.1)^2 + (38 - 34.1)^2 = 53.23$

We also know the degrees of freedom for between-group variability ($df_{between}$) = 2 (because there are three categories for our independent variable (top, middle, sides) and we subtract 1.

Now we can find between-group variability.

$$\text{Between-group variability} = \frac{SS_{between}}{df_{between}} = MS_{between} = \frac{53.23}{2} = 26.62$$

Within-group variability:

By knowing each sample's sum-of-squares, we can find SS_{within}.

$$s_T{}^2 = \frac{\Sigma(x_{Ti} - \bar{x}_T)^2}{n_T - 1} \xrightarrow{therefore} \Sigma(x_{Ti} - \bar{x}_T)^2 = (s_T{}^2)(n_T - 1)$$

$$s_M{}^2 = \frac{\Sigma(x_{Mi} - \bar{x}_M)^2}{n_M - 1} \xrightarrow{therefore} \Sigma(x_{Mi} - \bar{x}_M)^2 = (s_M{}^2)(n_M - 1)$$

$$s_S{}^2 = \frac{\Sigma(x_{Si} - \bar{x}_S)^2}{n_S - 1} \xrightarrow{therefore} \Sigma(x_{Si} - \bar{x}_S)^2 = (s_S{}^2)(n_S - 1)$$

$SS_{within} = (s_T{}^2)(n_T - 1) + (s_M{}^2)(n_M - 1) + (s_S{}^2)(n_S - 1)$

$= 3.2^2(235 - 1) + 4.3^2(169 - 1) + 2.5^2(210 - 1) = 6808.73$

We also know that the degrees of freedom for within-group variability ($df_{between}$) equals the total number of values minus the number of categories.

N – k = 235 + 169 + 210 – 3 = 611

Finally, we can find within-group variability (i.e. MS_{within}) by dividing SS_{within} by df_{within}.

Within-group variability = $\frac{SS_{within}}{df_{within}} = MS_{within}$

$$= \frac{6808.73}{611} = 11.14$$

F-statistic:

Finally, we can find the F-statistic.

F = (between-group variability) / (within-group variability) = $\frac{26.62}{11.14}$ = 2.39

We can organize all our calculations in an ANOVA table. You'll get the same output when you do an ANOVA test in R, but the SS and df columns will be switched.

Table 7: Results of between-group and within-group variability calculations with F-statistic finding.

	SS	df	MS	F
Page location	53.23	2	26.62	2.39
Error	6808.73	611	11.14	

Once we've calculated the F statistic, we can compare this to the F-critical value $F_{(0.05, 2, 611)} = 3$.

Because the F-statistic is less than $F_{(0.05, 2, 611)}$, the results are not significant. Therefore, we fail to reject the null, and we conclude there is no evidence to suggest any two of the populations—where each population is the CTR for each ad location for all publications—are significantly different. In other words, there is no significant difference between the CTRs of different ad locations on web pages.

Now that we've performed one-way ANOVA by hand, let's execute it in R with the NCES data. Let's say we want to know if SES differs significantly by race (where subjects are coded 0 if they are "White"—see the Appendix for labels corresponding to other races). We'll first apply the **`tapply()`** function, which tells us the statistic (e.g., mean, median) we will specify for a particular variable as broken out by the other variable. The **`tapply(ses, race, mean)`** function will take the mean of SES as broken out by race.

Next, we'll use the term "fit" to name the ANOVA test, then we'll summarize "fit" to see the results.

Code Listing 11

```
> tapply(ses, race, mean) #calculates the mean of income2011 based on the
variable "race"

          0           1           2           3           4
 0.26154596 -0.29982456  0.06211096 -0.15877778 -0.39424116
          5           6
-0.17442561  0.09525253

> fit = aov(ses ~ as.factor(race)) #label "fit" as the ANOVA test, ensure
that the variable "race" is treated as a factor using as.factor()
function

> summary(fit) #summarize results of the ANOVA test

                  Df Sum Sq Mean Sq F value Pr(>F)
as.factor(race)    6    354   58.94     119 <2e-16 ***
Residuals       8240   4080    0.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can see from the **`tapply()`** function that white students (code = 0) are of higher SES than students of other races. The ANOVA test confirms that these differences are significant because the F-statistic is huge (119). The presence of asterisks (*) also signifies significance.

After determining that at least two populations are significantly different, we can determine which two populations by running a **post hoc test**. Choosing which post hoc test best suits our needs will be dependent upon our initial data. If our original data has about the same variance (i.e. homogeneity of variance), we can use a test called Tukey's HSD (for "honestly significant difference"). If not, we can use the Games Howell test. First, let's go through a test to determine whether or not variances are homogenous.

Levene's test for homogeneity of variance

In this test, the null and alternative hypotheses are:

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$
H_a: at least two population variances are significantly different.

How the statistic is calculated goes beyond the scope of this e-book; however, you'll learn how to do this in R.

Because R is open source, people all over the world develop packages containing new functions that perform various statistical tests. For example, the "lawstat" package contains the **`levene.test()`** function, and you can install this package in order to run this test for homogeneity of variance.

First, download the package from cran.r-project.org/web/packages/lawstat/. The following code listing explains how to install the package in R and how to run Levene's test using the ANOVA example of how SES differs by race.

Code Listing 12

```
> install.packages("lawstat") #installs the package into R

--- Please select a CRAN mirror for use in this session ---
#select the location closest to your current location

> row.names(installed.packages()) #ensure that "lawstat" is listed as one
of the packages currently installed

> library(lawstat) #load the packages into R

> levene.test(ses, race) #run Levene's Test for homogeneity of variance

      modified robust Brown-Forsythe Levene-type test based
      on the absolute deviations from the median

data:  ses
Test Statistic = 19.4668, p-value < 2.2e-16
```

Tip: If you're not sure how to use a particular function in R, you can type a ? followed by the function name, and R will output information on each input of the function. For example, typing `?levene.test()` and hitting 'enter' will show you how to use this function.

According to this test, variances are not homogenous. Therefore, we would use the Games Howell test to determine which students have significantly different SES by race. First, let's pretend the results of Levene's test were not significant and go over how we would perform Tukey's HSD in R.

Tukey's HSD (variances are homogenous)

The **`TukeyHSD()`** test outputs the absolute differences between the means of each group (in this case, the difference between the mean SES), the lower and upper bounds of the 95% confidence interval for these differences, and the p-value (where p-values less than α indicate honestly significant differences).

Code Listing 13

```
> TukeyHSD(fit)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ses ~ as.factor(race))

$`as.factor(race)`
           diff         lwr         upr     p adj
1-0 -0.56137052 -0.83770881 -0.28503223 0.0000000
2-0 -0.19943500 -0.28097831 -0.11789169 0.0000000
3-0 -0.42032374 -0.49867291 -0.34197456 0.0000000
4-0 -0.65578712 -0.75465341 -0.55692083 0.0000000
5-0 -0.43597157 -0.53047888 -0.34146427 0.0000000
6-0 -0.16629343 -0.27444148 -0.05814539 0.0001191
2-1  0.36193552  0.07668733  0.64718372 0.0034779
3-1  0.14104678 -0.14330479  0.42539835 0.7669639
4-1 -0.09441660 -0.38509128  0.19625807 0.9627701
5-1  0.12539895 -0.16382216  0.41462006 0.8619244
6-1  0.39507709  0.10111583  0.68903835 0.0014499
3-2 -0.22088874 -0.32644573 -0.11533175 0.0000000
4-2 -0.45635212 -0.57791786 -0.33478639 0.0000000
5-2 -0.23653657 -0.35458451 -0.11848864 0.0000001
6-2  0.03314156 -0.09608570  0.16236883 0.9888601
4-3 -0.23546339 -0.35491007 -0.11601670 0.0000001
5-3 -0.01564783 -0.13151240  0.10021673 0.9996921
6-3  0.25403030  0.12679443  0.38126618 0.0000001
5-4  0.21981555  0.08919952  0.35043158 0.0000146
6-4  0.48949369  0.34869271  0.63029467 0.0000000
6-5  0.26967814  0.13190294  0.40745333 0.0000002
```

You can see from the results of the Tukey's HSD test that the only races that do not have significantly different SES levels are between:

- Groups 3 and 1 ("Black or African American, non-Hispanic" and "Amer. Indian/Alaska Native, non-Hispanic").
- Groups 4 and 1 ("Hispanic, no race specified" and "Amer. Indian/Alaska Native, non-Hispanic").

- Groups 5 and 1 (“Hispanic, race specified” and “Amer. Indian/Alaska Native, non-Hispanic”).
- Groups 6 and 2 (“More than one race, non-Hispanic” and “Asian, Hawaii/Pac. Islander, non-Hispanic”).
- Groups 5 and 3 (“Hispanic, race specified” and “Black or African American, non-Hispanic”).

However, remember that the data did not pass our homogeneity of variance test, so let’s see what the Games Howell test has to say.

Games Howell (variances are not homogenous)

A package called “userfriendlyscience”[2] includes a convenient **`oneway()`** function that can output the results of a number of tests we specify in addition to the ANOVA results, including Levene’s test and the Games Howell test. Let’s run the **`oneway()`** function in R and include results of the Levene’s test to compare with the results we got earlier from the “lawstat” package.

Code Listing 14

```
> install.packages("userfriendlyscience") #installs the package into R

--- Please select a CRAN mirror for use in this session ---
#select the location closest to your current location

> library(userfriendlyscience) #load the packages into R

> oneway(ses, as.factor(race), posthoc="games-howell", levene=TRUE) #this
will run ANOVA for the variable “ses” by the factor “race”, and
additionally perform the Games Howell post hoc test and Levene’s test

### Oneway Anova for y=ses and x=as.factor(race) (groups: 0, 1, 2, 3, 4,
5, 6)

Eta Squared: 95% CI = [0.07; 0.09], point estimate = 0.08

                                    SS   Df    MS      F
Between groups (error + effect)  353.64    6 58.94 119.05
Within groups (error only)      4079.57 8240   0.5
                                    p
Between groups (error + effect) <.001
Within groups (error only)
```

[2] Information about the functions included in the “userfriendlyscience” package can be found at http://cran.r-project.org/web/packages/userfriendlyscience/userfriendlyscience.pdf.

```
### Levene's test:

Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)
group    6  19.864 < 2.2e-16 ***
      8240
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

### Post hoc test: games-howell

        t      df     p
0:1  6.24   57.23 <.001
0:2  6.14  876.58 <.001
0:3 16.02 1058.23 <.001
0:4 19.25  561.13 <.001
0:5 12.40  613.47 <.001
0:6  4.91  462.84 <.001
1:2  3.82   70.29  .005
1:3  1.52   64.71  .731
1:4  0.99   71.89  .954
1:5  1.31   73.09  .845
1:6  4.15   71.62  .002
2:3  5.58 1434.84 <.001
2:4 10.10 1135.28 <.001
2:5  5.14 1191.55 <.001
2:6  0.74 1000.13  .990
3:4  5.75  984.95 <.001
3:5  0.37 1042.74 1.000
3:6  6.24  839.26 <.001
4:5  4.66 1009.76 <.001
4:6 10.60  867.93 <.001
5:6  5.74  914.77 <.001
```

We get the same results for ANOVA and Levene’s test using the **`oneway()`** function. We know from ANOVA that at least two groups are significantly different (i.e. at least two races have different SES levels), and we know from Levene’s test that there is not homogeneity of variance in SES between the different races.

Finally, it turns out that the Games Howell test gives us the same results as Tukey’s HSD.

Compare samples based on two factors

You just learned one-way ANOVA, which compares groups based on one independent variable (i.e. one factor). In our example, that factor was the location of the ad on the web page (top, middle, or sides). The dependent variable was the ad’s click-through rate.

Now let's work with **two-way ANOVA**, which tests for significant differences based on two factors. For example, you may want to see how 2011 income differs by gender and whether or not getting good grades as a student differentiates income. Not only are you interested in the difference in income based on gender and grades, but also in a possible **interaction effect** between gender and grades (i.e. is the difference in income consistent between genders irrespective of grades received and is the difference in income consistent between grades irrespective of gender?).

There are three null hypotheses in two-way ANOVA:

- The means for Factor 1 are equal.
- The means for Factor 2 are equal.
- There is no interaction effect between Factor 1 and Factor 2.

Two-way ANOVA outputs an F-statistic for each hypothesis that determines whether we reject or fail to reject it. Calculating these statistics by hand gets very complicated, so we'll simply review the basic principles before performing the analysis in R.

Each factor has a certain number of categories (e.g., in the NCES data, the factor "gender" has two categories: "male" and "female"; the factor "grades" has two categories: "Yes" and "No" in regard to whether or not the student was recognized for good grades). Let's say Factor 1 has k categories and Factor 2 has q categories.

Each category contains a certain number of numeric values. Let's use n to represent the number of values in Factor 1, where n_1 is the number of values in Category 1 of Factor 1, n_2 is the number of values in Category 2 of Factor 1, etc., through n_k, which is the number of values in Category k of Factor 1.

Let's use m to represent the number of values in Factor 2, where m_1 is the number of values in Category 1 of Factor 2, m_2 is the number of values in Category 2 of Factor 2, etc., through n_q, which is the number of values in Category q of Factor 2.

Table 8: Two-way ANOVA has two factors (i.e. independent variables). Let's say Factor 1 has k categories and n total values, and Factor 2 has q categories and m total values.

Factor 1	Category 1 Category 2 … Category k	n_1 values n_2 values … n_k values
Factor 2	Category 1 Category 2 … Category q	m_1 values m_2 values … m_q values

Our goal in two-way ANOVA is the same as with one-way ANOVA: we want to calculate measures of between-group variability and divide by a measure of error (within-group variability) to calculate our F-statistic. However, this time we'll calculate two additional F-statistics: one for the second factor and one for the interaction between the two factors. We symbolize this interaction with a multiplication sign.

Let's first calculate the sums-of-squares for Factor 1 and Factor 2 ($SS_{between}$ for each), the interaction of Factors 1 and 2 ($SS_{1\times2}$), the error (SS_{within}), and SS_{total}.

We can make these calculations easier by organizing the means of each category in a table.

Table 9: The Mean Table organizes both the mean of each subset based on each factor and the marginal means (the mean of all values in each category of each factor). Marginal means are used to calculate between-group variability, while the means of each bucket (in a specific category of each factor) are used to calculate within-group variability.

Mean Table

		Factor 1 (k categories)			
		Category 1 n_1 values	**Category 2 n_2 values**	**Category 3 n_3 values**	**Marginal Mean**
Factor 2 (q categories)	Category 1 m_1 values	$\bar{x}_{1,1}$	$\bar{x}_{2,1}$	$\bar{x}_{3,1}$	$\bar{x}_{i,1}$ = sum of all values in Factor 2 Category 1 divided by m_1
	Category 2 m_2 values	$\bar{x}_{1,2}$	$\bar{x}_{2,2}$	$\bar{x}_{3,2}$	$\bar{x}_{i,2}$ = sum of all values in Factor 2 Category 2 divided by m_2
	Marginal Mean	$\bar{x}_{1,i}$ = sum of all values in Factor 1 Category 1 divided by n_1	$\bar{x}_{2,i}$ = sum of all values in Factor 1 Category 2 divided by n_2	$\bar{x}_{3,i}$ = sum of all values in Factor 1 Category 3 divided by n_3	$\bar{x}_G$= sum of all values in data set divided by total number of values (N)

$\bar{x}_G$ can also be found by taking the weighted averages of the averages of each category, i.e. by choosing one of the factors, multiplying the average of each category by the sample size, and dividing the sum by N, where $N = n_1 + n_2 + n_3 = m_1 + m_2$.

$$\bar{x}_G = \frac{n_1(\bar{x}_{1,j}) + n_2(\bar{x}_{2,j}) + n_3(\bar{x}_{3,j})}{n_1 + n_2 + n_3} = \frac{m_1(\bar{x}_{i,1}) + m_2(\bar{x}_{i,2})}{m_1 + m_2}$$

You can see we have six buckets of values associated with one of the categories from Factor 1 and one of the categories from Factor 2.

Now we can calculate the sums-of-squares for Factor 1 and Factor 2. SS_1 is found by subtracting the grand mean $\bar{x}_G$ from the mean of each category from Factor 1, squaring each deviation, multiplying each squared deviation by the sample size for that category, and taking the sum. Similarly, SS_2 is found by subtracting $\bar{x}_G$ from the mean of each category from Factor 2, etc.

$$SS_1 = n_1(\bar{x}_{1,j} - \bar{x}_G)^2 + n_2(\bar{x}_{2,j} - \bar{x}_G)^2 + \cdots + n_k(\bar{x}_{k,j} - \bar{x}_G)^2$$

$$SS_2 = m_1(\bar{x}_{i,1} - \bar{x}_G)^2 + m_2(\bar{x}_{i,2} - \bar{x}_G)^2 + \cdots + m_q(\bar{x}_{i,q} - \bar{x}_G)^2$$

Note: These equations assume there are k categories for Factor 1 and q categories for Factor 2; however, Table 1 shows that k = 3 and q = 2.

To calculate the sum-of-squares for within-group variability (our error term), subtract the mean of each bucket (Factor 1 Category i and Factor 2 Category j) from each value in that bucket, square each deviation, and take the sum.

$$SS_{within} = \Sigma(x_{h,1,1} - \bar{x}_{1,1})^2 + \Sigma(x_{h,2,1} - \bar{x}_{2,1})^2 + \cdots + \Sigma(x_{h,k,1} - \bar{x}_{k,1})^2$$
$$+ \Sigma(x_{h,1,2} - \bar{x}_{1,2})^2 + \Sigma(x_{h,2,2} - \bar{x}_{2,2})^2 + \cdots + \Sigma(x_{h,q,2} - \bar{x}_{q,2})^2$$

Note: Here's a translation—this equation takes the first value in Factor 1, Category 1 and Factor 2, Category 1 (the top left box in the mean table) and subtracts the mean of all values in Factor 1, Category 1 and in Factor 2, Category 1, then continues this for every value in the data set. In other words, it sums the squared deviations of each value from the mean of all values in that same bucket (Factor 1 Category i and Factor 2 Category j). The above equation for SS_{within} uses h to represent the h'th value of each category ($x_{h,1,1}$ is the h'th value of Factor 1, Category 1 and Factor 2, Category 1).

The total sum-of-squares is the sum of the squared deviations of every value in the data set from the grand mean.

$$SS_{total} = \Sigma(x_h - \bar{x}_G)^2$$

Because all the sums-of-squares sum to SS_{total} ($SS_1 + SS_2 + SS_{1\times2} + SS_{within} = SS_{total}$), we can calculate the sum-of-squares of the interaction ($SS_{1\times2}$) by subtracting SS_1, SS_2, and SS_{within} from SS_{total}.

After finding each sum-of-squares, we need to know the degrees of freedom in order to calculate the mean square for each factor (MS_1 and MS_2), the error (MS_{within}), and the interaction ($MS_{1\times2}$). The F-statistics will be each mean square divided by MS_{within}.

Again, it helps to organize these calculations in a table.

Table 10: The ANOVA Table organizes the sums-of-squares (SS) for Factor 1, Factor 2, the interaction (Factor 1 × Factor 2), the error (within-group variability), and in total; the degrees of freedom (df); the mean squares (MS), which is SS/df; and the F-statistics (F), which tell us whether or not values of the dependent variable significantly differ by Factor 1, by Factor 2, or due to an interaction between the two.

	SS	**df**	**MS**	**F**
Factor 1	SS_1	$df_1 = k - 1$	$MS_1 = SS_1 / df_1$	MS_1/MS_{within}
Factor 2	SS_2	$df_2 = q - 1$	$MS2 = SS_2 / df_2$	MS_2/MS_{within}
Interaction	$SS_{1\times2}$	$df_1 \times df_2$	$MS_{1\times2} = (SS_1 \times SS_2)/(df_1 \times df_2)$	$MS_{1\times2} / MS_{within}$
Error	SS_{within}	$df_{within} = N - q \times k$	$MS_{within} = SS_{within}/df_{within}$	
Total	$SS_{total} =$ $SS_1 + SS_2 + SS_1 \times SS_2 + SS_{within}$	$df_{total} =$ $df_1 + df_2 + df_1 \times df_2 + df_{within} =$ $N - 1$		

This should give you an idea of how we find each F-statistic, which, like one-way ANOVA, is the ratio of between-group variability to within-group variability.

Let's now perform ANOVA on the NCES data using R with "income2011" as the dependent variable and "gender" and "grades" as the two factors or independent variables. We have coded subjects who are male as 0 and subjects who are female as 1. Subjects with bad grades are coded as 0 and subjects with good grades as 1.

Code Listing 15

```
> tapply(income2011, gender, mean) #find the mean of income2011 by each
gender

       0        1
30968.73 24013.91

> tapply(income2011, grades, mean) #find the mean of income2011 by
whether or not the student was recognized for good grades

       0        1
```

```
23995.63 30166.17

> income = aov(income2011 ~ as.factor(gender)*as.factor(grades)) #test
whether or not gender, grades, and the interaction between gender and
grades are significant

> summary(income) #return the F statistics

                                     Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(gender)                     1 9.943e+10 9.943e+10 172.127 <2e-16 ***
as.factor(grades)                     1 9.928e+10 9.928e+10 171.875 <2e-16 ***
as.factor(gender):as.factor(grades)   1 2.600e+09 2.600e+09   4.501 0.0339 *
Residuals                          8243 4.762e+12 5.777e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can make a few observations using the **`tapply()`** function: males made an average of about $7,000 more than females, and students recognized for good grades made more than $6,000 annually in 2011.

When we perform ANOVA, we see that not only does income significantly differ by gender and whether or not students had good grades—there is also an interaction between these two factors, indicating that income is not consistent between grades when you separate students by gender. We can better understand this interaction effect by finding the mean income for each of the four groups (males who received good grades, females who received good grades, males who did not receive good grades, and females who did not receive good grades).

Code Listing 16

```
> tapply(income2011[grades=="0"], gender[grades=="0"], mean) #find the
mean income in 2011 for students who did not receive good grades,
separated by gender (again, for those who did not receive good grades)

       0        1
28179.41 19272.01

> tapply(income2011[grades=="1"], gender[grades=="1"], mean) #find the
mean income in 2011 for students who received good grades, separated by
gender (again, for those who received good grades)

       0        1
33998.32 27357.42
```

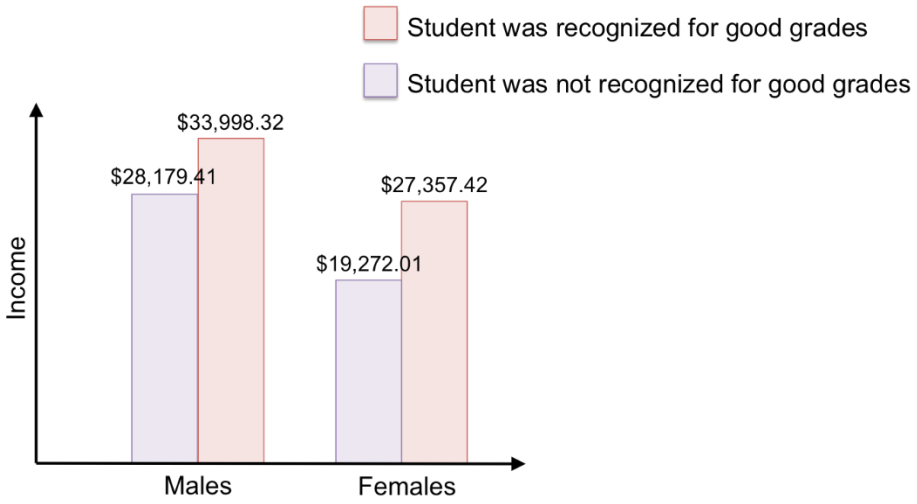


Figure 40: This bar graph visualizes the mean income for each group. Now we can see the interaction between gender and grades more clearly: females had a larger difference in income due to grades than did males.

Additional ANOVA information

Make sure you understand the data before accepting the results of ANOVA. For example, does your data have outliers? These can result in a Type I error (rejecting the null hypothesis when it is, in fact, true) or Type II error (failing to reject the null when there is, in fact, a significant difference) by pulling the mean toward it, so that the mean is not a good measure of center.

For ANOVA to have accurate results, the dependent variable should be normally distributed, have roughly equal sample sizes, and have roughly equal variances. Of course, this is not often the case, but there are methods to transform your data so that it no longer violates these assumptions. These additional tests and transformation methods are beyond the scope of our material, but there is a plethora of information online that describes what you can do for your data's situation.

In the next lesson, you'll learn statistical tests for tabulated data—determining whether or not frequencies (rather than specific values) differ significantly from what was expected.

Chapter 8 Tabulated Data

Test for significance with tabulated data

You've learned how to determine whether or not two or more samples comprised of continuous data are significantly different. We can also perform hypothesis tests for tabulated data (a tally of subjects that fits into various categories) to determine if proportions significantly differ and whether or not the number of values in subsets of data significantly differs from the expected number of values. The former involves a z-test; the latter involves a chi-square test.

Difference between proportions

The z-test test is similar to the test for proportions you learned in Chapter 5. However, this time we're comparing two samples rather than comparing one proportion to an expected proportion. The null and alternative hypotheses are:

H_0: $p_1 = p_2$
H_a: $p_1 \neq p_2$

For example, let's say you want to know if divorce is more likely to occur among urban or suburban professionals. You send out surveys to randomly selected professionals aged 30-50 in major cities and various suburbs across the U.S. asking if they've ever been divorced (response is "yes" or "no"). You get back 1032 responses from urban professionals and 865 responses from suburban professionals. The results are tabulated in Table 11.

Table 11: Survey results for use in a z-test.

	Have you ever been divorced?	
	Yes	**No**
Urban	187	845
Suburban	62	803

You can use a z-test to determine if these results are significant and if one group has higher divorce rates than the other.

As we learned in Chapter 7, we need to calculate a z-score by finding the difference between the two proportions (the proportion of urban professionals who responded one way and the proportion of suburban professionals who also responded that way) and divide this difference by the standard error.

Let's analyze the proportion that responded "yes." For urban professionals, $p_1 = 187/1032 = 0.18$. For suburban professionals, the proportion that responded "yes" is $p_2 = 62/865 = 0.07$. So we'll look at the difference between 0.18 and 0.07 and divide this difference by the standard error.

In this case, the standard error changes because we need to account for the two sample proportions as well as the two sample sizes. To do this, we calculate a pooled sample proportion, $\hat{p}$.

$$\hat{p} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

We then use $\hat{p}$ to calculate the standard error, SE.

$$\text{SE} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Now, we can calculate our z-statistic.

$$z = \frac{p_1 - p_2}{SE}$$

Let's perform this z-test for proportions with our example, starting with calculating the pooled sample proportion.

$$\hat{p} = \frac{(0.18)(1032) + (0.07)(865)}{1032 + 865} = 0.13$$

We can now calculate the standard error.

$$\text{SE} = \sqrt{(0.13)(1-0.13)\left(\frac{1}{1032} + \frac{1}{865}\right)} = 0.016$$

Finally, we can calculate the z-score.

$$z = \frac{0.18 - 0.07}{0.016} = 6.875$$

Because this z-score is far greater than 1.96, the z-critical value for a two-tailed test at $\alpha = 0.05$, we'll reject the null and can conclude that the two proportions are significantly different—urban professionals are more likely to get divorced.

Chi-square test

We can analyze this same tabulated data with a **chi-square test**, which is different from the z-test in that it compares the frequencies of occurance to what we might expect if the two factors are independent, i.e. we can't predict the level of one factor by knowing the other.

H_0: The two factors are independent.
H_a: The two factors are not independent (we can predict the frequency of one factor by knowing that of the other).

Table 12: Survey results for use in a chi-square test.

	Have you ever been divorced?		
	Yes	**No**	**Total**
Urban	187	845	1032
Suburban	62	803	865
Total	249	1648	1897

In this case, we would expect that the number of urbanites who have been divorced is the same proportion of the total number who have been divorced (249/1897 = 0.13). So, if about 13% of all people have been divorced and we have 1032 urbanite responses, we would expect that 0.13×1032 = 134.16 urbanites have been divorced. Looking at our data, we see a higher number of divorces (187) than this expected value (134). We want to determine if this difference is significant. Let's first go a little more in-depth with how we find the expected values.

To simplify the procedure for finding expected values, we multiply the marginal totals and divide by the grand total (1,897).

Table 13: Expected values (in green) are found by multiplying each marginal total and dividing by the grand total.

	Have you ever been divorced?		
	Yes	**No**	**Total**
Urban	187 $\frac{(249)(1032)}{1897} = 135.46$	845 $\frac{(1648)(1032)}{1897} = 896.54$	1032
Suburban	62 $\frac{(249)(865)}{1897} = 113.54$	803 $\frac{(1648)(865)}{1897} = 751.46$	865
Total	249	1648	1897

After calculating expected values, we compute a **chi-square (χ^2)** statistic

$$\chi^2 = \Sigma \frac{(f_O - f_E)^2}{f_E}$$

where f_0 is the observed value and f_E is the expected value. Here is our example:

$$\chi^2 = \frac{(187 - 135.46)^2}{135.46} + \frac{(845 - 896.54)^2}{896.54} + \frac{(62 - 113.54)^2}{113.54} + \frac{(803 - 751.46)^2}{751.46} = 49.5$$

Again, we use another table to determine if our results (i.e. the difference between our observed and expected values) are significant. Degrees of freedom are equal to (n-1)(m-1), where n is the number of categories for Factor 1 and m is the number of categories for Factor 2. In this case, there are two categories for location (urban and suburban) and two categories for divorce (yes and no). Therefore, df = (2-1)(2-1) = 1. The chi-square table tells us that for df = 1 and $\alpha = 0.05$, the critical χ^2 value is 3.84. Because our computed χ^2 statistic is greater than the critical value, we conclude that location (urban vs. suburban) and whether or not someone has been divorced are independent of one another.

Chapter 9 Linear Regression

Predict one variable with another

We've looked at tests that determine whether or not two or more groups of values are significantly different or independent. With ANOVA, you can determine whether or not the dependent variable is significantly different between values of the categorical, independent variable(s).

The remainder of this e-book addresses how to test the association between one or more independent variables and a dependent variable—in other words, how to predict the amount by which the dependent variable will change when an independent variable changes by x, with all else held constant. This allows us to extrapolate—make future predictions based on trends—and interpolate—estimate values of one variable based on values of another. For example, given how the world population has been changing over the last 10 years, what might we expect the world population to be in the year 2050 (extrapolation)? Or, given the relationship between house price and square feet in a certain location, what might we expect to be the price of a 2,000-square-foot home (interpolation)?

We do this by finding a model that fits the trends in the data, one that inputs specified values of the independent variable(s) and outputs the predicted value of the dependent variable. This test is called **regression**, and the model we derive is the **regression model**. While there are many types of regression models (e.g., logistic, quadratic), we will cover only linear models, which are the simplest.

Because regression looks at the change in the dependent variable associated with a change in the independent variable(s), both the independent and dependent variables must be numeric.

Correlation

When performing linear regression, we first visualize trends in the data with a scatter plot. (Note that we can only do this with one independent variable.) Scatter plots show values of the independent variable on the x-axis (horizontal axis) and values of the dependent variable on the y-axis (vertical axis). For this reason, we'll call the independent variable "x" and the dependent variable "y." This visualization can quickly tell us whether or not the relationship between x and y is strong (points form a straighter line), weak (points are more variable), positive (greater values of the independent variable are associated with greater values of the dependent variable), or negative (greater values of the independent variable are associated with smaller values of the dependent variable).

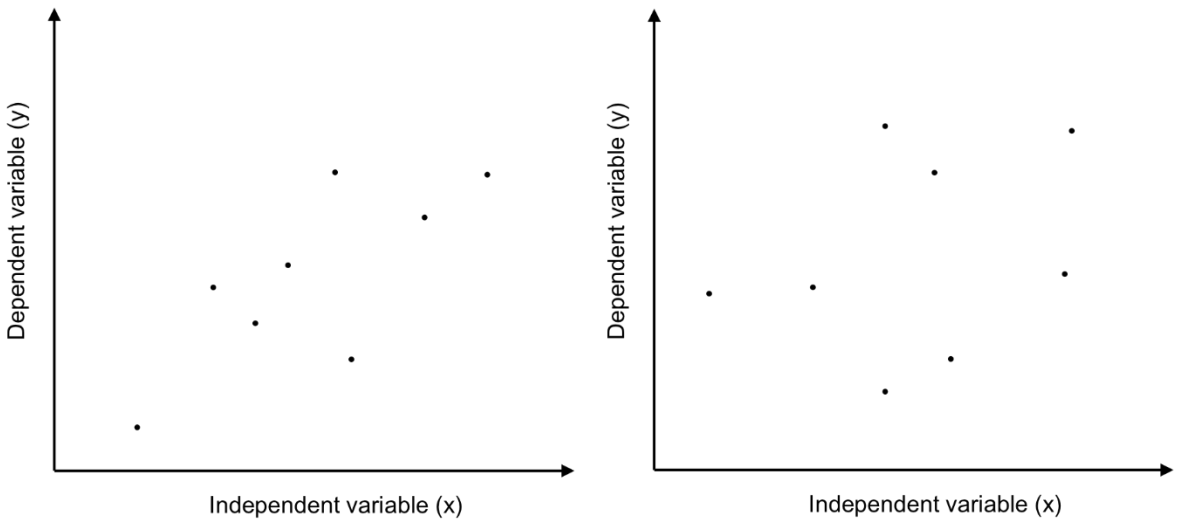


Figure 41: The scatter plot on the left shows a positive, strong relationship; the scatter plot on the right shows a positive, weak relationship.

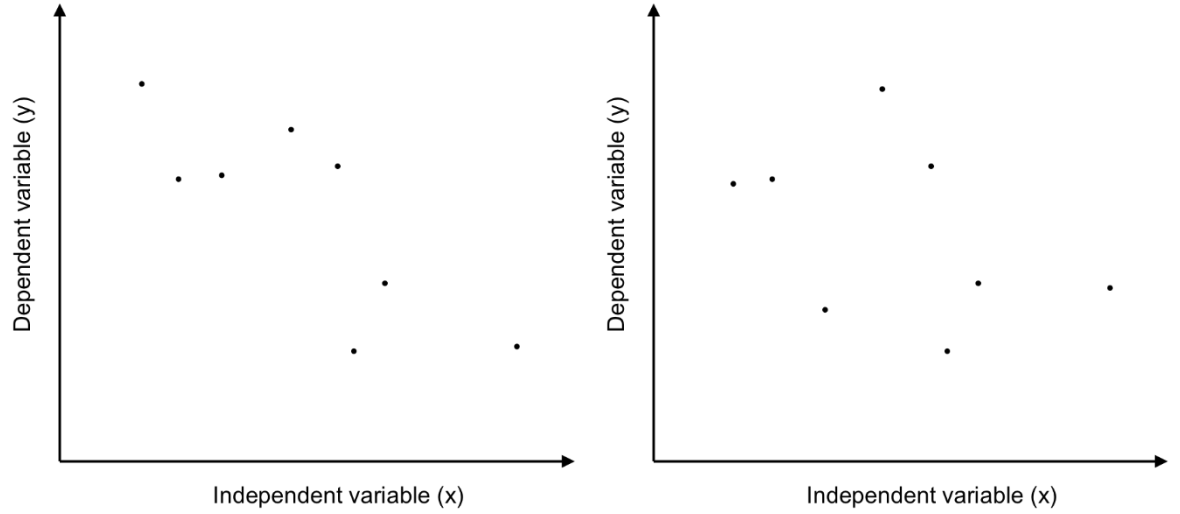


Figure 42: The scatter plot on the left shows a negative, strong relationship; the scatter plot on the right shows a negative, weak relationship.

We can quantify the strength and direction of a relationship with a statistic called the **correlation coefficient**, which is denoted by r. While the sign of r indicates the direction, the distance r is from 0 indicates the relationship's strength.

Note that r ranges from -1 to 1, where -1 is a perfect negative relationship (i.e. the points form a straight line), and 1 is a perfect positive relationship. When r = 0, there is no relationship between x and y.

To calculate r, we first find the **covariance** of x and y, which measures the association that a change in x has with a change in y, by finding the average product of each x and y value from their respective means. If there is no relationship between x and y, then some of these products will be negative and some will be positive, and they will cancel each other out, resulting in a covariance closer to 0.

$$cov_{x,y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

You can see from this equation that a positive relationship between x and y means that most points (x_i, y_i) are to the lower left and the upper right of $(\bar{x}, \bar{y})$, so that $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are usually either both positive or both negative. This will result in the products being mostly positive (so the sum $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ will also be positive). Similarly, a negative relationship between x and y means that most coordinates are to the upper left and lower right of $(\bar{x}, \bar{y})$, resulting in a negative covariance.

To find r, we divide the covariance by the product of the standard deviation of x and the standard deviation of y.

$$r = \frac{cov_{x,y}}{(s_x)(s_y)}$$

Note: Because the standard deviation is always positive, the covariance determines whether r is positive or negative and therefore is the statistic responsible for describing the direction.

The product $(s_x)(s_y)$ will always be greater than or equal to $cov_{x,y}$. If you visualize each, you can see that $(s_x)(s_y)$ is the product of squares, which maximize area, and $cov_{x,y}$ is the product of rectangles.

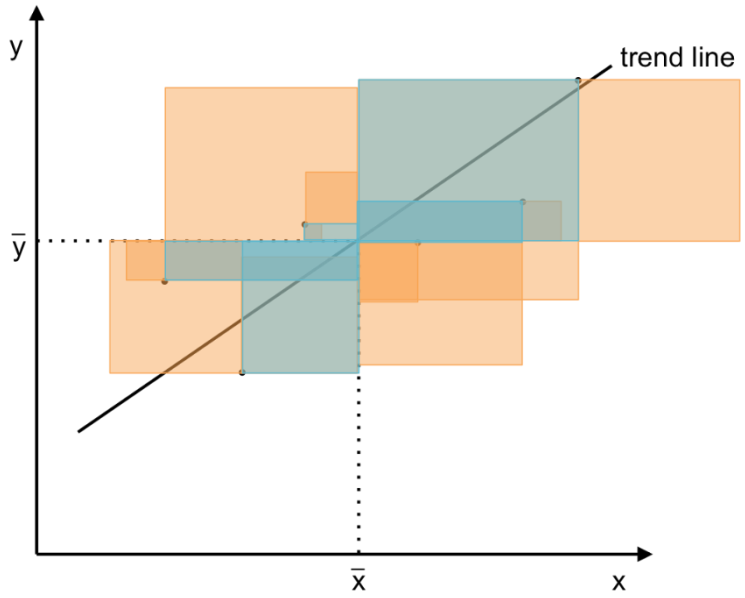


Figure 43: The covariance is the area of the average blue rectangle, while $(s_x)(s_y)$ is the standard length of the orange squares multiplied by the standard height of the orange squares (where "standard" is the square root of the area of the average orange square).

When r equals 1 or -1, the covariance is equal to the product of the standard deviations (r = 1) or the negative product of the standard deviations (r = -1).

As in previous chapters, when we calculate the correlation, we want to do a hypothesis test for significance. This test helps us decide—based on our calculation of r—if the true correlation of the population (denoted ρ) is significantly different from 0.

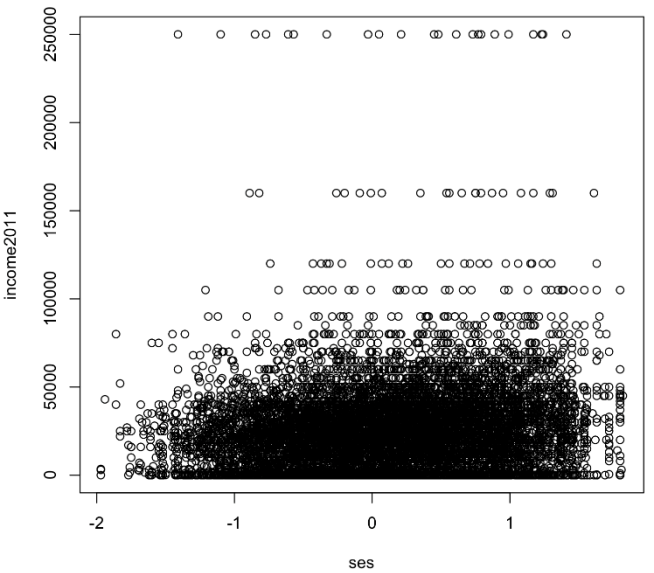
H0: $\rho = 0$
Ha: $\rho \neq 0$ (two-tailed test)

Again, this is a type of t-test. We will not address how to calculate the t-statistic; the important thing is that you understand the principles and can interpret the results.

Let's do a correlation test in R between SES and income in 2011 from the NCES data.

Code Listing 17

```
> plot(ses, income2011) #creates a scatter plot with "ses" on the x-axis
and "income2011" on the y-axis
```




```
> cor.test(ses, income2011)

      Pearson's product-moment correlation

data:  ses and income2011
t = 13.4525, df = 8245, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1253655 0.1676048
sample estimates:
      cor
0.1465519
```

You can see from this test that while r is small (0.15), the results are significant, meaning that we're pretty sure ρ is significantly different from 0. Also note that R gives us the 95% confidence interval for ρ, the entire range of which is positive.

Line of best fit

After determining that a relationship does indeed exist between the independent and dependent variables, the next step is to predict how much the dependent variable will change when one or more of the independent variables changes by a certain amount. We do this using a **regression line**, or **line of best fit**, so named because it minimizes the sum-of-squared **residuals** (the distance between each observed y-value (y_i) and the predicted value ($\hat{y}_i$) for the corresponding observed x-value (x_i)). The sum-of-squared residuals is equal to $\Sigma(y_i - \hat{y}_i)^2$. We'll first go through **simple linear regression**, which involves only one independent variable.

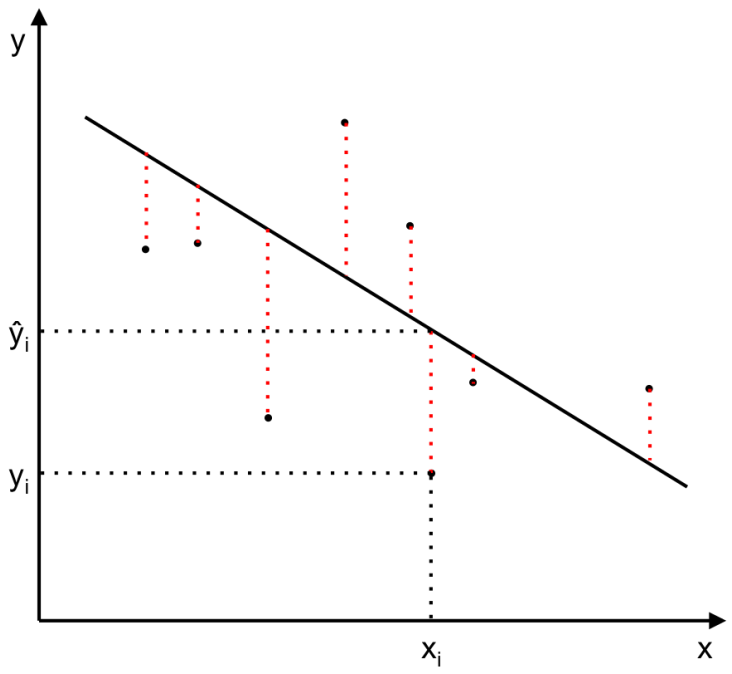


Figure 44: The red dotted lines visualize the risiduals (y_i - $\hat{y}_i$). Each point visualizes the observed values (x_i, y_i), and the line of best fit shows the predicted values of y ($\hat{y}$) for any value of x. This line minimizes the sum-of-squared residuals.

The general equation for the regression line is $\hat{y}_i = b_0 + b_1 x_i$, where b_0 and b_1 are called the regression coefficients. Coefficient b_0 is the predicted value of y when x = 0; coefficient b_1 is the amount by which y is expected to change when x changes by one unit.

We can determine the values of b_0 and b_1 by using calculus to minimize $\Sigma(y_i - \hat{y}_i)^2$, setting $\hat{y}_i$ equal to $b_0 + b_1 x_i$, inputting each value of x_i and y_i, and knowing that the regression line will always pass through the point $(\bar{x}, \bar{y})$.

$$b_1 = r\left(\frac{s_x}{s_y}\right)$$

$$b_0 = \bar{y} - r\left(\frac{s_x}{s_y}\right)\bar{x}$$

Therefore, this is our linear regression equation:

$$\hat{y}_i = \bar{y} - r\left(\frac{s_x}{s_y}\right)\bar{x} + r\left(\frac{s_x}{s_y}\right)x_i$$

We won't bother calculating this by hand; instead, we'll do it in R. And R will perform another hypothesis test to determine if the slope b_1 is significantly different than 0. In other words, we want to know if a change in x is indeed associated with a change in y.

Let's first execute a linear regression analysis with income2011 as the dependent variable and only SES as the independent variable. Then we'll perform a multiple regression analysis to predict how a change in multiple independent variables would lead to a change in income2011.

Code Listing 18

```
> lm.income1 = lm(income2011 ~ ses) #assigns the name lm.income1 to the
regression analysis

> summary(lm.income1) #outputs the results of the regression analysis

Call:
lm(formula = income2011 ~ ses)

Residuals:
   Min     1Q Median     3Q    Max
-35519 -16468  -2624  10514 230221

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26693.0      271.1   98.48   <2e-16 ***
ses           4903.4      364.5   13.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24270 on 8245 degrees of freedom
Multiple R-squared:  0.02148, Adjusted R-squared:  0.02136
F-statistic:   181 on 1 and 8245 DF,  p-value: < 2.2e-16
```

The results of this test show that the coefficient of SES is 4,903.4, meaning that an increase in SES of 1 is associated with an increased 2011 salary of $4,903.4. This is significant at $\alpha <$ 0.001, meaning that the true population likely has the same association between SES and 2011 income.

Note: Because the units of SES don't contain much meaning, it's helpful to analyze the mean, standard deviation, range, and distribution so that we can see what a one-unit increase in SES means. If we create a histogram of SES, we can see that the increase is approximately normally distributed with mean 0.12 and standard deviation 0.73. We can then find the z-score of 1.12 (the

mean plus an increase in SES of 1), which is 1.34. Therefore, a one-unit increase in SES moves from average to the top 90%.

With **multiple regression**, we have n independent variables, and our general model is $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$. Each coefficient $b_0 \ldots b_n$ tells us how much y is expected to change when the corresponding independent variable changes by one unit. Our null hypothesis states that each true coefficient for the population B_0, B_1, … B_n, is equal to 0 (a change in the respective independent variable is not associated with a change in the dependent variable) and the alternative states that the coefficient is significantly different from 0.

Let's do an example in R. We can use the **`cor.test()`** function to find that the standardized test score ("test") has a significant correlation with income in 2011 ($r = 0.17$, $p < 0.001$). Perhaps we want to predict 2011 income with the test score as well as with demographic variables race, gender, and SES.

Code Listing 19

```
> lm.income2 = lm(income2011 ~ test + race + gender + ses) #assigns the
name lm.income2 to the regression analysis

> summary(lm.income2) #outputs the results of the regression analysis

Call:
lm(formula = income2011 ~ test + race + gender + ses)

Residuals:
   Min     1Q Median     3Q    Max
-42251 -15636  -2389  10599 237684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13357.78    1717.06   7.779 8.17e-15 ***
test          332.02      31.26  10.621  < 2e-16 ***
race         -285.83     141.05  -2.026   0.0427 *
gender      -6626.44     526.92 -12.576  < 2e-16 ***
ses          2612.20     404.27   6.462 1.10e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23860 on 8242 degrees of freedom
Multiple R-squared:  0.05444, Adjusted R-squared:  0.05399
F-statistic: 118.6 on 4 and 8242 DF,  p-value: < 2.2e-16
```

The results of this test show that all coefficients are significant—"test," "gender," and "ses" at $p < 0.001$ and "race" at $p < 0.05$. The associations "test" and "ses" have with "income2011" are easy to interpret because they're numeric and continuous—a one-unit increase in test score is associated with a $332.02 increase in 2011 salary (with all else constant), and a one-unit increase in SES is associated with a $2,612.20 increase in salary. Note that the coefficient for SES was $4,903 when we executed simple linear regression. It changes now because we're factoring in other variables.

The coefficients for "race" and "gender" are a little trickier because these variables are categorical. For "race," we mostly care about the sign (positive or negative) and the reference value ("White," because we assigned it a value of 0). This means any increase in this variable (i.e. moving from White to non-White) is associated with a smaller 2011 income.

Because "male" is the reference value for the variable "gender" ("male" is coded 0, "female" is coded 1), moving from "male" to "female" is associated with a decrease of $6,626.44 in 2011 income.

Feel free to test other multiple regressions using independent variables, such as the relationship between standardized test score ("test") and variables such as whether or not students played sports, watched television, got good grades, etc. Perhaps you'll find some quantitative evidence for which behaviors lead to good grades, which may come in handy when convincing a recalcitrant student to study!

Afterward

Continue your statistics journey

We've examined several methods for describing and analyzing data, and hopefully you now understand when to use these tests and how to perform them. Of course, the tests presented are only several of hundreds of statistical tests you can perform, but nevertheless, remember these two important takeaways:

- **The way we think about data sets is universal.** We should always understand the data intimately by knowing descriptive statistics—mean, median, mode, minimum, maximum, range, standard deviation, variance—and the distribution.

  Additionally, you should understand where the data comes from and how it was gathered. Take the NCES data, for example. Recall that the sample size only consisted of students who responded to all questions. Researchers must analyze who did and did not respond to certain questions and factor this in when using samples to draw conclusions about a population. Did a certain demographic (e.g., students of higher SES) respond to or avoid particular questions? If so, the resulting sample size may have **nonresponse bias**.

- **We should determine which test(s) we perform based on the data we have.** These tests are like different models of cars. We need to consider our needs in order to choose a car that will meet our requirements.

As you encounter different sets of data, you should now have the vocabulary and foundational understanding to ask the questions that will lead you to the findings you seek.

Glossary

Alpha level (α): Also known as the significance level, the alpha level defines the probability for which you consider a value or sample to be unlikely to occur. It is equal to the area of the critical region. If $p < \alpha$, you reject the null hypothesis and conclude that your results are significant.

Alternative hypothesis (H_a): The alternative hypothesis states that results are significant (in the case of a z- or t-test, it states that a significant difference exists between two particular values or samples).

Analysis of Variance (ANOVA): We use ANOVA to test whether or not any pair of three or more samples (which are continuous) differs significantly from each other as a result of different values of the independent variable (which is categorical). Two-way ANOVA additionally tells us whether or not an interaction effect exists between the two factors used in the analysis.

Box plot: A box plot visualizes the spread of data by showing where the minimum value, maximum value, Q_1, Q_2 (the median), Q_3, and any outliers are in relation to each other.

Central Limit Theorem: This theorem states that for a population with any distribution shape, the sampling distribution (the distribution of means of all possible samples of size n) will be normal with mean μ and standard deviation $\sigma / \sqrt{n}$.

Central tendency: This is a term that describes where a group of numbers gather. It can be measured using the mean, median, or mode.

Chi-square statistic (χ^2): Used in a chi-square test, $\chi^2 = \Sigma \frac{(f_O - f_E)^2}{f_E}$ where f_O are the observed values and f_E are the expected values.

Chi-square test: This test is used for tabulated, categorical data and tells us whether or not observed values (i.e. frequency of subjects that fall in each category) differ significantly from expected values.

Confidence interval: A confidence interval is a numerical range in which C% of values lie. We can calculate confidence intervals by knowing the percentage of values or sample means that fall within a certain number of standard deviations (z*) from the population mean μ.

Correlation coefficient (r): The correlation coefficient is a single statistic that describes the strength and direction of the relationship between two variables, where r ranges from -1 (perfect negative relationship) and 1 (perfect positive relationship)—with values closer to 0 indicating a weak relationship.

Covariance ($cov_{x,y}$): The covariance measures the association that a change in x has with a change in y by finding the average product of each x and y value from their respective means.

$$cov_{x,y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Critical region: This is the area on a distribution in which a value or sample is considered significantly unlikely. If the sample statistic is farther on the tail of a distribution than the critical value (which marks the cutoff of the critical region), results are significant. The probability of landing in the critical region is equal to α.

Data: Data is what we use to draw all our statistical conclusions. Before doing any analysis, we must first understand the type of data (e.g., categorical vs. numerical, discrete vs. continuous), visualize it (e.g., by graphing a histogram), and describe it (at the very least, find the minimum, maximum, mean, median, and standard deviation).

Degrees of freedom (df): These are the number of values that can change given specified parameters. For example, if the mean is known, the first n-1 values can be anything, but the last value must be such that the mean remains what was specified. Therefore, in this example, n-1 is the degrees of freedom.

Dependent variable: This variable is usually thought to be influenced by the independent variable (though the tests in this e-book are for association rather than causality).

Deviation: This is the difference between a particular value and the expected value (often the mean).

Distribution: A distribution is the shape created when graphing the frequency of values in a data set in particular ranges or bins. Common distributions include normal, uniform, skewed, and bimodal.

Expected values: These are the values we would expect to obtain based on one or more factors. In the case of one sample, if we randomly select a particular value, the expected value is the mean. In regression, expected values of y (the dependent variable) are based on the value of x (the independent variable).

F-critical value ($F_{\alpha, k-1, N-k}$): This is the cutoff of the critical region used in ANOVA. Here, k - 1 is the degrees of freedom for between-group variability (one less the number of samples, k) while N - k is the degrees of freedom for within-group variability (k less than the total number of values, N).

Frequency: This is the number of values in a particular category or range.

Histogram: Histograms visualize the frequency of values in ordered, numerical bins so that we can easily see how the data is distributed.

Hypothesis testing: This procedure is used in many statistical tests to determine whether or not a value, sample, proportion, etc., is significantly different than expected. It involves stating the null and alternative hypotheses and comparing a statistic (a measure of the difference between observed and expected values) to a critical value to determine if that statistic is significantly large. If the statistic is greater than the critical value, results are considered significant.

Independent variable: This variable is usually thought to influence the dependent variable (though the tests in this e-book are for association rather than causality).

Interquartile range (IQR): This is a measure of spread that ignores extreme values by finding the difference between the third and first quartiles ($Q_3 – Q_1$). The rectangles in box plots visualize the IQR.

Mean: The mean is a measure of center that takes every value in a data set into account in its calculation. The arithmetic mean sums all values in the sample or population and (respectively) divides by the sample size (n) or population size (N).

Measure of center: This is a measure of central tendency. Three of the most common measures of center are the mean, median, and mode. Each contributes to understanding a data set in its own way; one is not always a more accurate measure than another.

Measure of spread: This is a measure of variability. The range (the difference between the maximum and minimum values) is one of the simplest measures of spread. Other measures include the IQR, variance, average absolute deviation, and the more commonly used standard deviation.

Median: The median is a measure of center representing the 50^{th} percentile, i.e. half the values in the data set are less than the median and half are greater. For odd-numbered data sets, the median is the middle value, and for even-numbered data sets, the median is the average of the two middle values.

Mode: The mode is a measure of center representing the value (in the case of discrete data), range of values (in the case of large data sets or continuous data), or category (in the case of categorical data) with the greatest frequency.

Null hypothesis (H_0): In hypothesis testing, the null hypothesis represents the outcome that results are not significant.

Observed values: These are the actual values found in the sample. Observed values most often do not follow a perfect pattern, which means we need to measure the error, or variability.

Outliers: These are values in a data set that differ significantly from other values; specifically, they are less than 1.5 IQRs below Q_1 or more than 1.5 IQRs above Q_3.

Population: The population consists of everyone or everything with the characteristic(s) under study rather than a subset of them.

Post hoc test: This is a test performed after ANOVA in order to determine which two groups are significantly different. Tukey's HSD can be used when variances are roughly equal; the Games Howell test can be used when they're not. Levene's Test for Homogeneity of Variance helps us decide which test to perform.

Probability density function (PDF): This curve models the distribution of a data set. The area under the PDF is equal to the cumulative probability (thus, the total area under the PDF is 1).

Range: The range is a measure of spread equal to the maximum value minus the minimum value.

Regression: This statistical procedure involves modeling the trend between a dependent variable and one or more independent variables, allowing you to calculate expected values of the dependent variable based on specified values of each independent variable.

Regression line: In linear regression, the regression line ($\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$) is an equation that outputs expected y-values ($\hat{y}$) based on observed x-values. It is also known as the line of best fit because it minimizes the sum of squared residuals $\Sigma(y_i - \hat{y}_i)^2$.

Residual: In regression, the residual is the difference between the observed (y_i) and expected value ($\hat{y}_i$) for a given observed x-value (x_i) equal to $y_i - \hat{y}_i$.

Sample: A sample is a subset of a population. For robust statistical analyses, samples should be randomly selected.

Sample standard deviation (s): When using a sample to estimate the population standard deviation, we divide the sum of squared deviations by n-1 (rather than n) to correct for the likely smaller standard deviation of the sample as compared to the population.

$$\sigma \approx s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Sampling distribution: This is the distribution of sample means from all samples of size n that can be taken from a population. The mean is equal to the population mean (μ) and the standard deviation is equal to the population standard deviation divided by the square root of the sample size ($\sigma\sqrt{}$).

Standard deviation (σ): This is the most common measure of spread used in statistical analyses. It is equal to the square root of the average squared deviation (the square root of the variance).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Standard error (SE): This is a measure of error often equal to the standard deviation of the distribution of interest, and it is the denominator in calculating statistics (e.g., the z, t, F) used to determine significance. For a z-score, the error is σ; for a sampling distribution, the error is $\sigma / \sqrt{n}$.

Standard normal distribution: The standard normal distribution, denoted N(0,1), has mean 0 and standard deviation 1. The z-table gives us cumulative probabilities under the standard normal curve for any z-score.

Standardize: This is the process of converting normal distributions into the standard normal curve in order to calculate probabilities.

Sum-of-squares (SS): This is the sum of squared deviations from the mean, e.g., $\Sigma(x_i - \bar{x})^2$.

t-critical value ($t_{\alpha, df}$): The t-critical value is the cutoff for which t-test results are significant. It is based on a chosen α level.

t-distribution: When we don't know the population standard deviation (σ) and so approximate it with the sample standard deviation (s), our results are prone to error. Therefore we use a t-distribution instead of a z-distribution. The fewer the degrees of freedom, the fatter the t-distribution's tails in order to account for the increase in error. As the degrees of freedom increases, the t-distribution better approximates the normal distribution.

t-statistic (t): The t-statistic is a measure that compares the difference between two values with the amount of error. If the difference is large enough or the error is small enough (i.e. the t-statistic lands farther on the tail of the t-distribution than the t-critical value), we reject the null.

t-test: We perform a t-test to determine significance when we don't know population parameters and we want to determine if a particular value from a relatively normal distribution is significant or if the difference between two samples is significant.

Variability: This is a term describing the error within a data set. Variance and standard deviation are both measures of variability.

Variance (σ^2): The variance is a measure of variability equal to the average squared deviation (i.e. the sum-of-squares divided by the sample size).

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

z-critical value (z*): The z-critical value defines the cutoff of the critical region defined by the chosen α level. If the z-score falls farther on the tails of the distribution than z*, the results are significant and we reject the null hypothesis.

z-score (z): The z-score indicates the number of standard deviations a value is from the mean. By knowing the z-score, we can find cumulative probabilities using the z-table.

z-test: We perform a z-test for significance when we know population parameters μ and σ. We can use a z-test to determine if a particular value or proportion falls significantly far from the mean of the population.

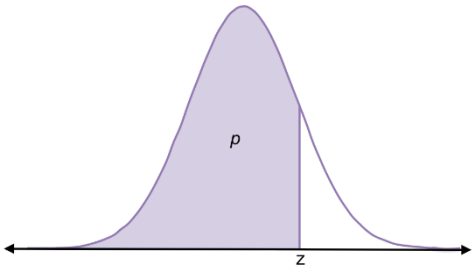
Appendix

Variables for NCES Education Longitudinal Study of 2002

Variable	Description	Values
gender	Student's gender.	0 = male 1 = female
race	Student's race.	0 = White, non-Hispanic 1 = Amer. Indian/Alaska Native, non-Hispanic 2 = Asian, Hawaii/Pac. Islander, non-Hispanic 3 = Black or African American, non-Hispanic 4 = Hispanic, no race specified 5 = Hispanic, race specified 6 = More than one race, non-Hispanic
ses	Measure of socioeconomic status based on five equally-weighted variables: father's/guardian's education, mother's/guardian's education, family income, prestige of father's/guardian's occupation, and prestige of mother's/guardian's occupation. Each of these five composite variables were inputed if missing.	Min: -1.97 Max: 1.82 Mean: 0.12 Median: 0.14 Standard deviation: 0.73
test	Average standardized test scores for math and reading.	Min: 20.91 Max: 81.04 Mean: 52.68 Median: 53.21 Standard deviation: 9.53
homework	Number of hours per week spent working on homework both in-school and out-of-school.	Min: 0 Max: 45 Mean: 9.85 Median: 8 Standard deviation: 7.46
tv_games	Average number of hours per weekday spent watching TV/videos or playing video/computer games.	Min: 0 Max: 8 Mean: 3.44 Median: 3 Standard deviation: 2.39
work	Student worked for pay during the school year.	0 = No 1 = Yes
grades	Student was recognized for good grades.	0 = No 1 = Yes
service	Student received community service award or participated in service club.	0 = No 1 = Yes
sports	Student participated in interscholastic sports at the junior varsity or varsity level.	0 = No 1 = Yes
music	Student participated in school band, chorus, and/or school play or musical.	0 = No 1 = Yes

student_gov	Student participated in student government.	0 = No 1 = Yes
honor	Student participated in academic honor society.	0 = No 1 = Yes
journalism	Student participated in school yearbook or newspaper.	0 = No 1 = Yes
vocation	Student participated in vocational education club or student organization.	0 = No 1 = Yes
income2011	2011 employment income.	Min: 0 Max: 250,000+ Mean: 27,302 Median: 24,000 Standard deviation: 24,532.78

Information about this data: *Data was downloaded from http://nces.ed.gov. Students with missing data were removed from the sample. n = 8247. To access the spreadsheet, visit http://turnthewheel.org/free-textbooks/street-smart-stats/afterward/ and click the first link under "Resources."*



z-table

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.496	0.492	0.488	0.484	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

t-table 1

Cum. Prob.	t(0.50)	t(0.75)	t(0.80)	t(0.85)	t(0.9)
one-tail	0.5	0.25	0.2	0.15	0.1
two-tails	1	0.5	0.4	0.3	0.2
df					
1	0	1	1.376	1.963	3.078
2	0	0.816	1.061	1.386	1.886
3	0	0.765	0.978	1.25	1.638
4	0	0.741	0.941	1.19	1.533
5	0	0.727	0.92	1.156	1.476
6	0	0.718	0.906	1.134	1.44
7	0	0.711	0.896	1.119	1.415
8	0	0.706	0.889	1.108	1.397
9	0	0.703	0.883	1.1	1.383
10	0	0.7	0.879	1.093	1.372
11	0	0.697	0.876	1.088	1.363
12	0	0.695	0.873	1.083	1.356
13	0	0.694	0.87	1.079	1.35
14	0	0.692	0.868	1.076	1.345
15	0	0.691	0.866	1.074	1.341
16	0	0.69	0.865	1.071	1.337
17	0	0.689	0.863	1.069	1.333
18	0	0.688	0.862	1.067	1.33
19	0	0.688	0.861	1.066	1.328
20	0	0.687	0.86	1.064	1.325
21	0	0.686	0.859	1.063	1.323
22	0	0.686	0.858	1.061	1.321
23	0	0.685	0.858	1.06	1.319
24	0	0.685	0.857	1.059	1.318
25	0	0.684	0.856	1.058	1.316
26	0	0.684	0.856	1.058	1.315
27	0	0.684	0.855	1.057	1.314
28	0	0.683	0.855	1.056	1.313
29	0	0.683	0.854	1.055	1.311
30	0	0.683	0.854	1.055	1.31
40	0	0.681	0.851	1.05	1.303
60	0	0.679	0.848	1.045	1.296
80	0	0.678	0.846	1.043	1.292
100	0	0.677	0.845	1.042	1.29
1000	0	0.675	0.842	1.037	1.282
z	0	0.674	0.842	1.036	1.282
	0%	50%	60%	70%	80%
	Confidence Level				

t-table 2

Cum. Prob.	t(0.975)	t(0.99)	t(0.995)	t(0.999)	t(0.9995)
one-tail	0.025	0.01	0.005	0.001	0.0005
two-tails	0.05	0.02	0.01	0.002	0.001
df					
1	12.71	31.82	63.66	318.31	636.62
2	4.303	6.965	9.925	22.327	31.599
3	3.182	4.541	5.841	10.215	12.924
4	2.776	3.747	4.604	7.173	8.61
5	2.571	3.365	4.032	5.893	6.869
6	2.447	3.143	3.707	5.208	5.959
7	2.365	2.998	3.499	4.785	5.408
8	2.306	2.896	3.355	4.501	5.041
9	2.262	2.821	3.25	4.297	4.781
10	2.228	2.764	3.169	4.144	4.587
11	2.201	2.718	3.106	4.025	4.437
12	2.179	2.681	3.055	3.93	4.318
13	2.16	2.65	3.012	3.852	4.221
14	2.145	2.624	2.977	3.787	4.14
15	2.131	2.602	2.947	3.733	4.073
16	2.12	2.583	2.921	3.686	4.015
17	2.11	2.567	2.898	3.646	3.965
18	2.101	2.552	2.878	3.61	3.922
19	2.093	2.539	2.861	3.579	3.883
20	2.086	2.528	2.845	3.552	3.85
21	2.08	2.518	2.831	3.527	3.819
22	2.074	2.508	2.819	3.505	3.792
23	2.069	2.5	2.807	3.485	3.768
24	2.064	2.492	2.797	3.467	3.745
25	2.06	2.485	2.787	3.45	3.725
26	2.056	2.479	2.779	3.435	3.707
27	2.052	2.473	2.771	3.421	3.69
28	2.048	2.467	2.763	3.408	3.674
29	2.045	2.462	2.756	3.396	3.659
30	2.042	2.457	2.75	3.385	3.646
40	2.021	2.423	2.704	3.307	3.551
60	2	2.39	2.66	3.232	3.46
80	1.99	2.374	2.639	3.195	3.416
100	1.984	2.364	2.626	3.174	3.39
1000	1.962	2.33	2.581	3.098	3.3
z	1.96	2.326	2.576	3.09	3.291
	95%	98%	99%	99.80%	99.90%
	Confidence Level				

f-table 1

/	df between = 1	2	3	4	5	6	7	8	9	10	12
df within = 1	161	200	216	226	230	236	237	239	241	242	244
2	18.5	19	19.1	19.3	19.3	19.3	19.3	19.4	19.5	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.79
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.6
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.53
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.31
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25
22	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34	2.3	2.23
23	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32	2.27	2.2
24	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25	2.18
25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24	2.16
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.2	2.13
28	4.2	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12
29	4.18	3.33	2.93	2.7	2.55	2.43	2.35	2.28	2.22	2.18	2.1
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2
60	4	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04	1.99	1.92
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83
∞	3.84	3	2.6	2.37	2.21	2.1	2.01	1.94	1.88	1.83	1.75

f-table 2

/	15	20	24	30	40	60	120	∞
df within = 1	246	248	249	250	251	252	253	254
2	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	8.7	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.86	5.8	5.77	5.75	5.72	5.69	5.66	5.63
5	4.62	4.56	4.53	4.5	4.46	4.43	4.4	4.37
6	3.94	3.87	3.84	3.81	3.77	3.74	3.7	3.67
7	3.51	3.44	3.41	3.38	3.34	3.3	3.27	3.23
8	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.01	2.94	2.9	2.86	2.83	2.79	2.75	2.71
10	2.85	2.77	2.74	2.7	2.66	2.62	2.58	2.54
11	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.4
12	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.3
13	2.53	2.46	2.42	2.38	2.34	2.3	2.25	2.21
14	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.4	2.33	2.29	2.25	2.2	2.16	2.11	2.07
16	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.31	2.23	2.19	2.15	2.1	2.06	2.01	1.96
18	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.2	2.12	2.08	2.04	1.99	1.95	1.9	1.84
21	2.18	2.1	2.05	2.01	1.96	1.92	1.87	1.81
22	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.07	1.99	1.95	1.9	1.85	1.8	1.75	1.69
27	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.03	1.94	1.9	1.85	1.81	1.75	1.7	1.64
30	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.84	1.75	1.7	1.65	1.59	1.53	1.47	1.39
120	1.75	1.66	1.61	1.55	1.5	1.43	1.35	1.25
∞	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1

χ^2 table 1

df	χ2 (.995)	χ2 (.990)	χ2 (.975)	χ2 (.950)	χ2 (.900)
1	0	0	0	0	0.02
2	0.01	0.02	0.05	0.1	0.21
3	0.07	0.12	0.22	0.35	0.58
4	0.21	0.3	0.48	0.71	1.06
5	0.41	0.55	0.83	1.15	1.61
6	0.68	0.87	1.24	1.64	2.2
7	0.99	1.24	1.69	2.17	2.83
8	1.34	1.65	2.18	2.73	3.49
9	1.74	2.09	2.7	3.33	4.17
10	2.16	2.56	3.25	3.94	4.87
11	2.6	3.05	3.82	4.58	5.58
12	3.07	3.57	4.4	5.23	6.3
13	3.57	4.11	5.01	5.89	7.04
14	4.08	4.66	5.63	6.57	7.79
15	4.6	5.23	6.26	7.26	8.55
16	5.14	5.81	6.91	7.96	9.31
17	5.7	6.41	7.56	8.67	10.09
18	6.27	7.02	8.23	9.39	10.87
19	6.84	7.63	8.91	10.12	11.65
20	7.43	8.26	9.59	10.85	12.44
21	8.03	8.9	10.28	11.59	13.24
22	8.64	9.54	10.98	12.34	14.04
23	9.26	10.2	11.69	13.09	14.85
24	9.89	10.86	12.4	13.85	15.66
25	10.52	11.52	13.12	14.61	16.47
26	11.16	12.2	13.84	15.38	17.29
27	11.81	12.88	14.57	16.15	18.11
28	12.46	13.57	15.31	16.93	18.94
29	13.12	14.26	16.05	17.71	19.77
30	13.79	14.95	16.79	18.49	20.6
40	20.71	22.16	24.43	26.51	29.05
50	27.99	29.71	32.36	34.76	37.69
60	35.53	37.49	40.48	43.19	46.46
70	43.28	45.44	48.76	51.74	55.33
80	51.17	53.54	57.15	60.39	64.28
90	59.2	61.75	65.65	69.13	73.29
100	67.33	70.07	74.22	77.93	82.36

χ^2 table 2

df	χ2 (.100)	χ2 (.050)	χ2 (.025)	χ2 (.010)	χ2 (.005)
1	2.71	3.84	5.02	6.64	7.88
2	4.61	5.99	7.38	9.21	10.6
3	6.25	7.82	9.35	11.35	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.65	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.54	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.73	26.76
12	18.55	21.03	23.34	26.22	28.3
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.69	26.12	29.14	31.32
15	22.31	25	27.49	30.58	32.8
16	23.54	26.3	28.85	32	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.2	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40
21	29.62	32.67	35.48	38.93	41.4
22	30.81	33.92	36.78	40.29	42.8
23	32.01	35.17	38.08	41.64	44.18
24	33.2	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.2	46.96	49.65
28	37.92	41.34	44.46	48.28	50.99
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67
40	51.81	55.76	59.34	63.69	66.77
50	63.17	67.51	71.42	76.15	79.49
60	74.4	79.08	83.3	88.38	91.95
70	85.53	90.53	95.02	100.43	104.22
80	96.58	101.88	106.63	112.33	116.32
90	107.57	113.15	118.14	124.12	128.3
100	118.5	124.34	129.56	135.81	140.17